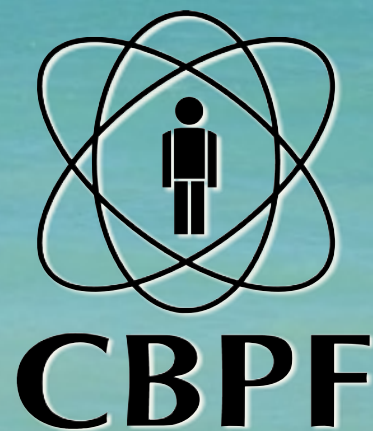
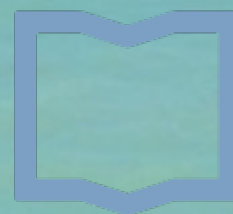


STATISTICAL TOOLS

MARTÍN MAKLER
CBPF & ICAS/IFICI/CONICET-UNSAM



MINISTRY OF
SCIENCE, TECHNOLOGY
AND INNOVATION



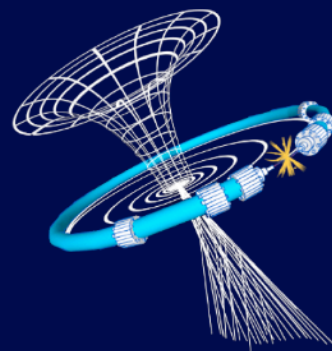
Instituto de
Ciencias Físicas
ICIFI-ECYT_UNSAM-CONICET



**15th IDPASC
Summer School**

20 - 29 July 2026

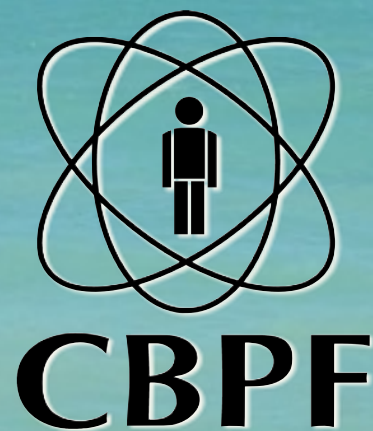
Natal, Brazil



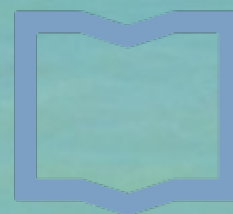
**INTERNATIONAL
INSTITUTE OF
PHYSICS**
Federal University of Rio Grande do Norte

SHORT INTRODUCTION TO A FEW STATISTICAL TOOLS

MARTÍN MAKLER
CBPF & ICAS/IFICI/CONICET-UNSAM



MINISTRY OF
SCIENCE, TECHNOLOGY
AND INNOVATION



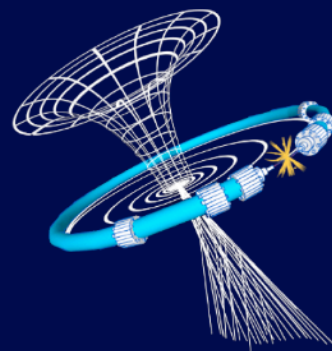
Instituto de
Ciencias Físicas
ICIFI-ECYT_UNSAM-CONICET



15th IDPASC
Summer School

20 - 29 July 2026

Natal, Brazil



INTERNATIONAL
INSTITUTE OF
PHYSICS

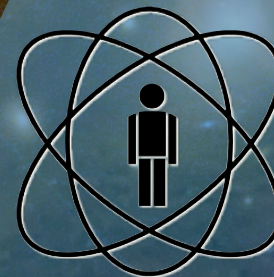
Federal University of Rio Grande do Norte

PASC@CBPF

COSMO:

10 researchers in astrophysics and
cosmology

COHEP: ~15 researchers in particle
physics and astrophysics

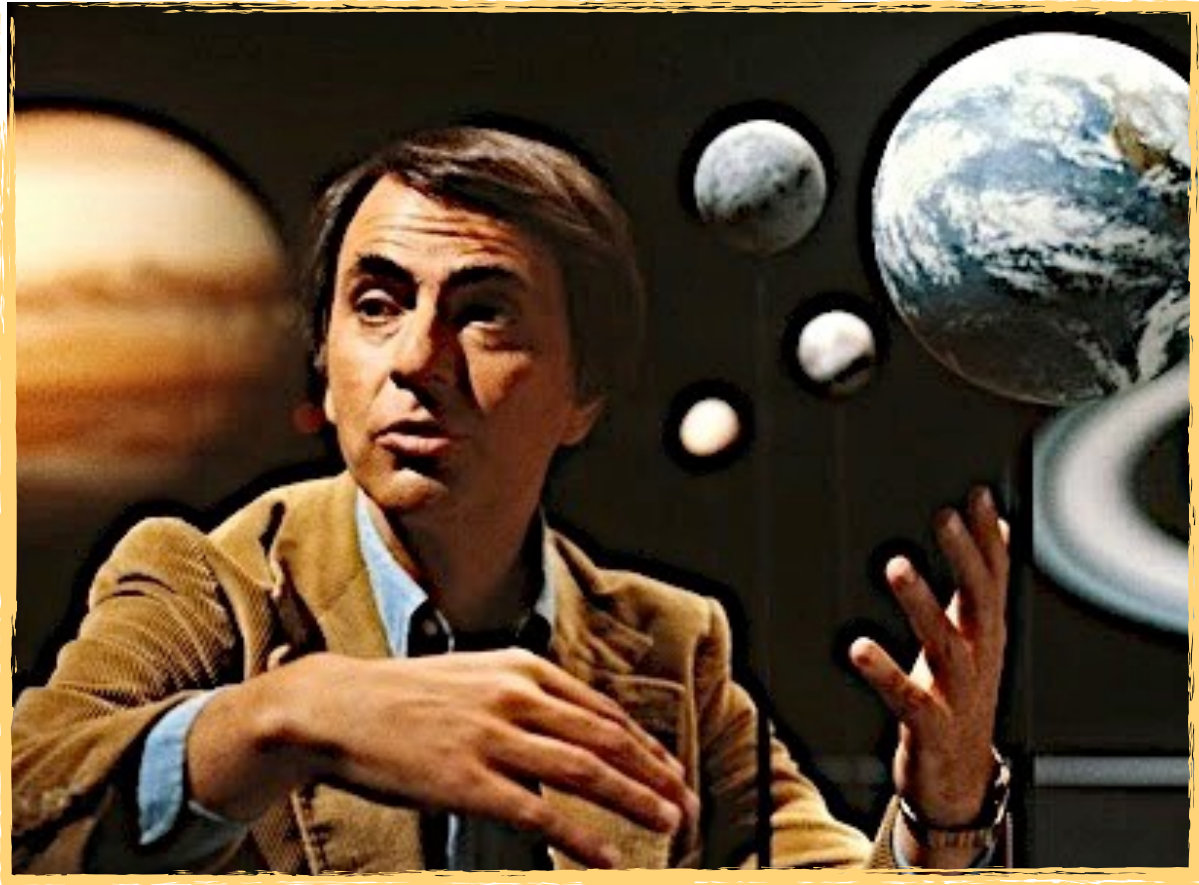


CBPF

MINISTRY OF
SCIENCE, TECHNOLOGY
AND INNOVATION



15TH IDEPASC



“Every question is a cry to understand the world. There is no such thing as a dumb question”

Carl Sagan

(The Demon-Haunted World: Science as a Candle in the Dark)

Introduction

- A word about the world cup
- A word about physics
 - Physics is about measurements/data/observation: need error bars (statistics!)
 - Physics is about connecting data with models: model fitting, errors, etc. (statistics!)
 - Vast field that underpins our daily practice

Plan

- Prolegomena on probability distributions

Part 1:

- Maximum Likelihood Estimation (MLE) and Likelihood Ratio confidence intervals
 - Worked examples

Part 2:

- A peek at Bayesian statistics
- On fitting, over (under) fitting, and regularization

A few mentis for fun



Join at menti.com | use code 4213 1500

Probability

- Determinism is more intuitive and acceptable than stochasticity
- Most decisions are based on a probabilistic assessment (conscious or not)
- Fundamental to our understanding of the world
- Not very intuitive
- Superficially addressed and poorly taught in most courses (including this one!)

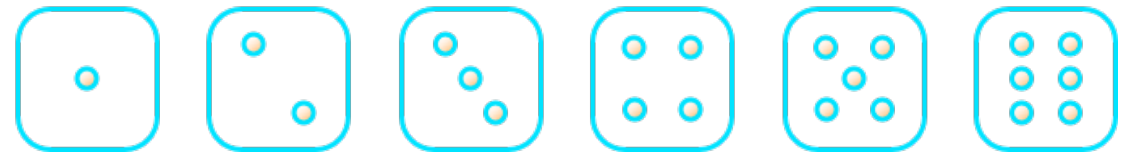
Probability

- Nomenclature: random, probabilistic, stochastic
- Sources of uncertainty
 - Incomplete observability
 - Incomplete modeling
 - Unpredictability (e.g. mood)
 - Inherent stochasticity (e.g. QM)
- Simple probabilities
 - Probability of something to happen (binary)
 - Integer probability (discrete)
 - Continuous probability

Probability distribution function

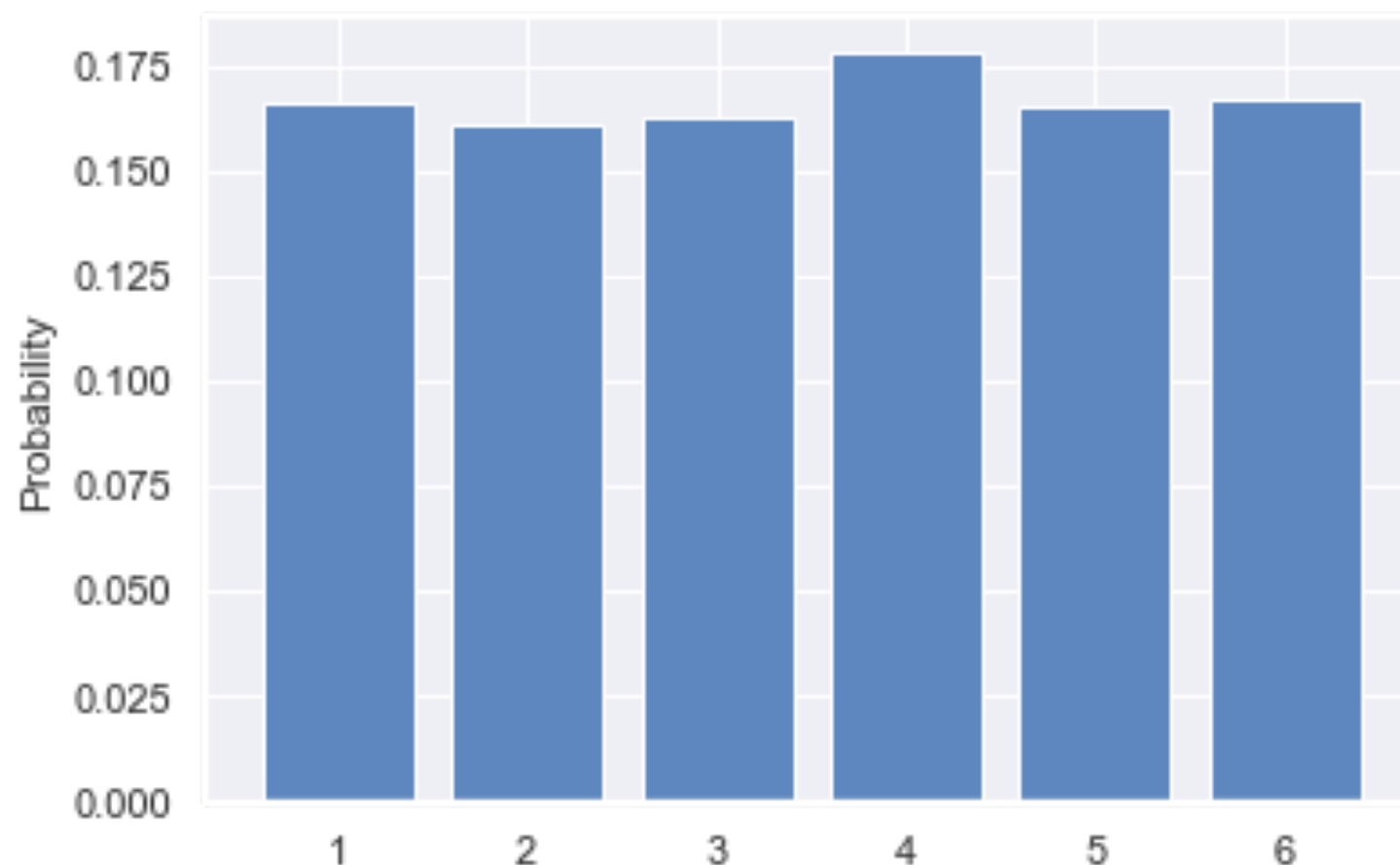
- Realization: outcome of a random process (e.g. roll a dice)

- Sampling: number of realizations



possible outcomes

- Example: 10,000 throws
- Outcome



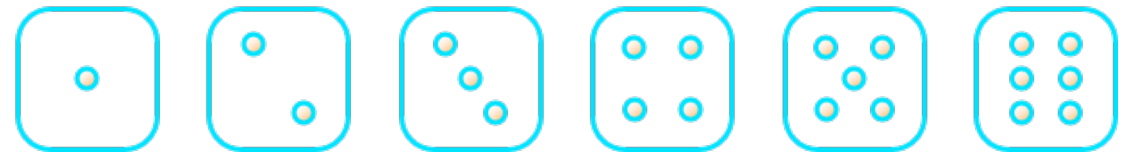
Relative **frequencies**:

$$\frac{\# \text{ times value appeared}}{\# \text{ number of throws}}$$

Probability distribution function

- Realization: outcome of a random process (e.g. roll a dice)

- Sampling: number of realizations

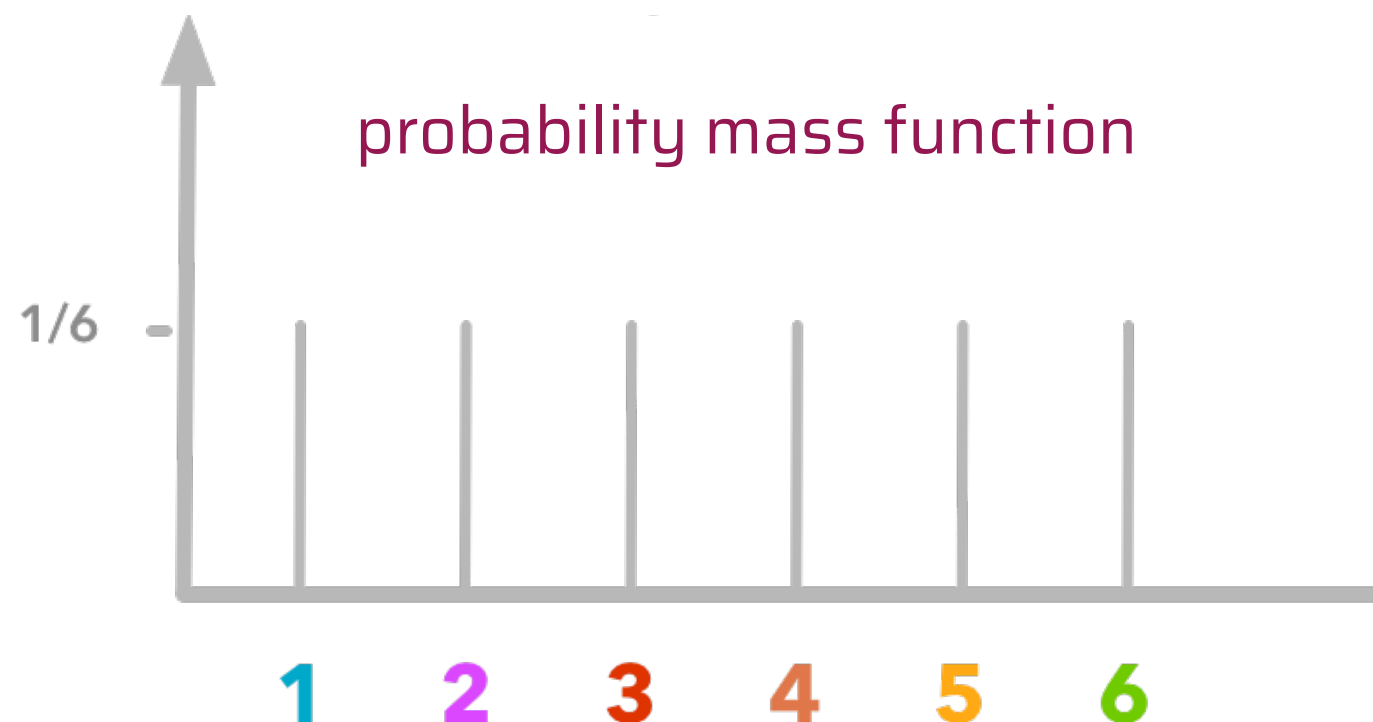


possible outcomes

- Example: 10,000 throws

- Outcome

Underlying probabilistic distribution

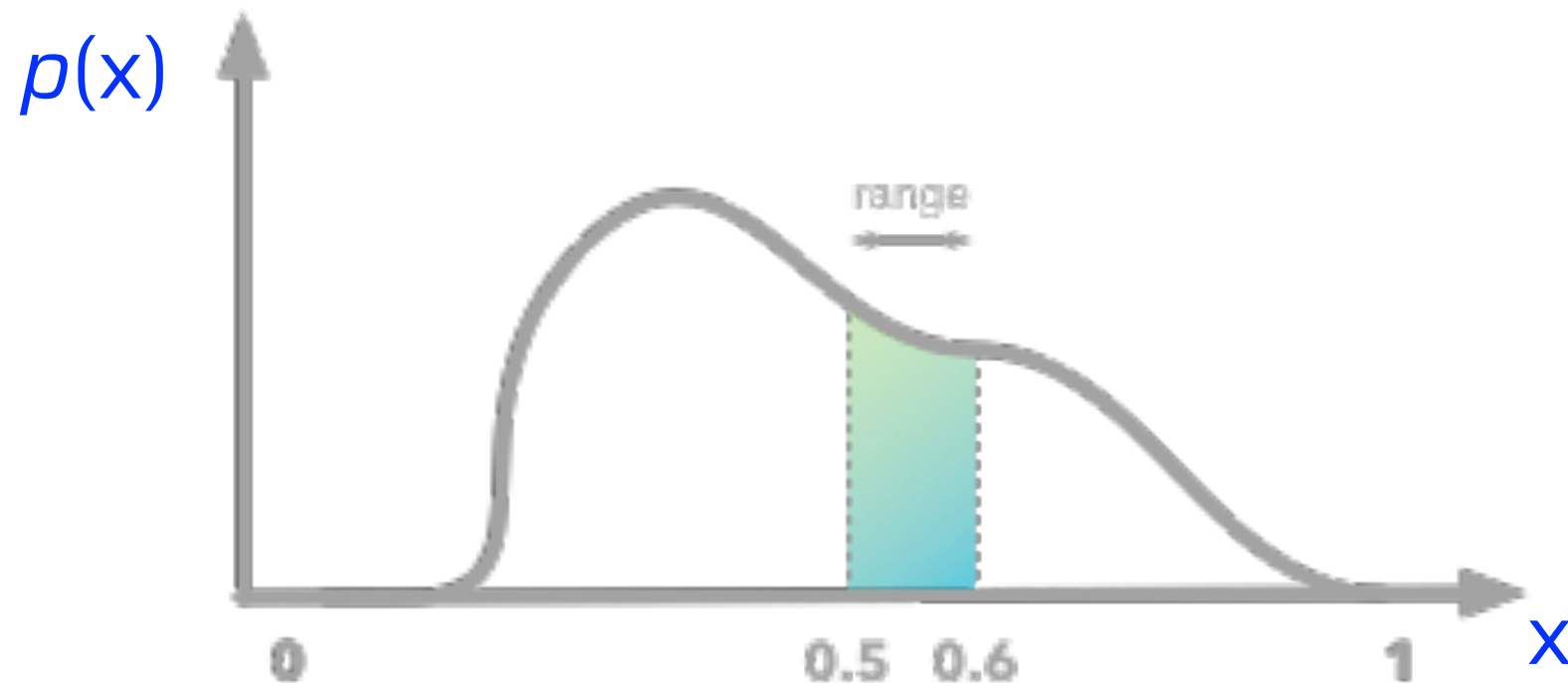


Probabilities add to one:

$$\sum_{x \in \mathbf{x}} P(x) = 1$$

Probability density/ distribution function

- A continuous variable will have continuous probability
- $p(x)$ is the probability density

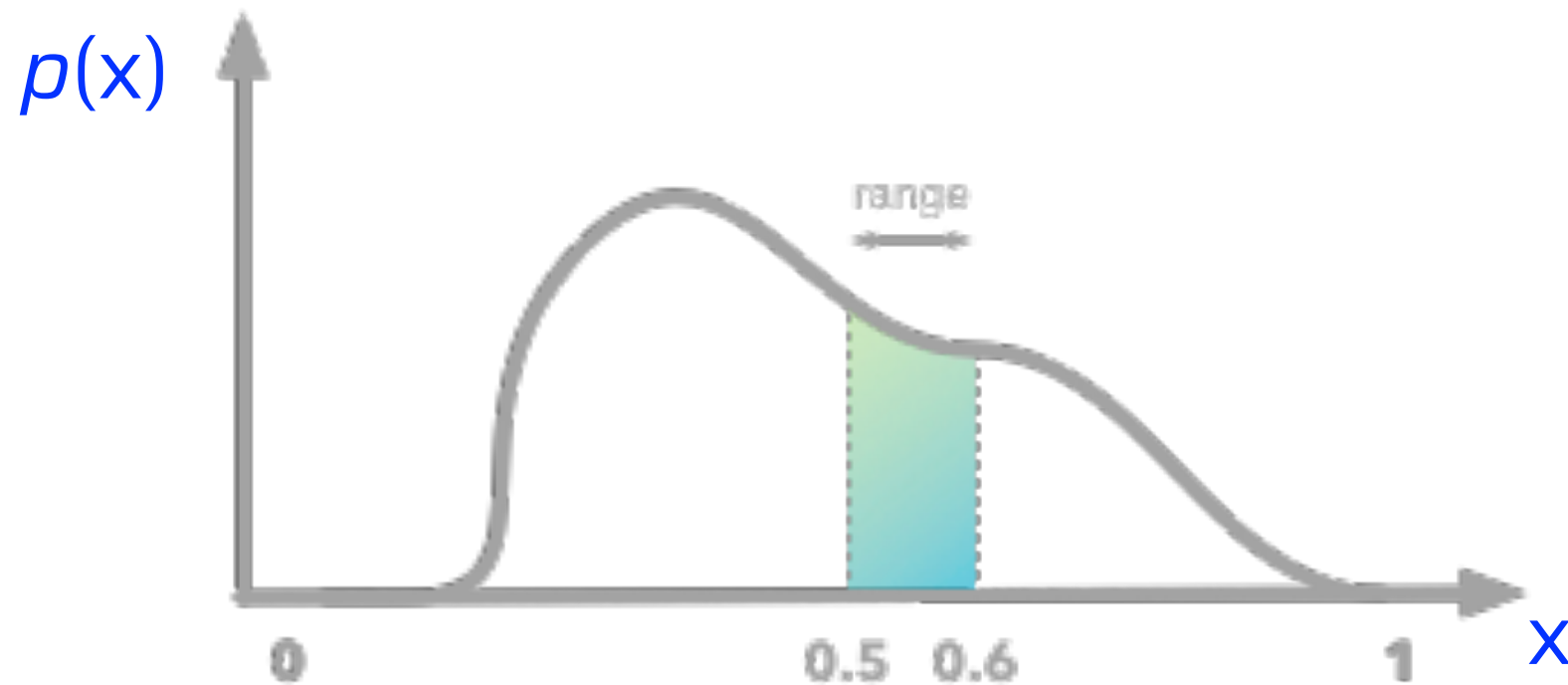


- Can compute a probability for a *range* of values

$$P = \int_A^B p(x) dx \quad \text{integral (area under the curve)}$$

Probability density/ distribution function

- A continuous variable will have continuous probability
- $p(x)$ is the probability density



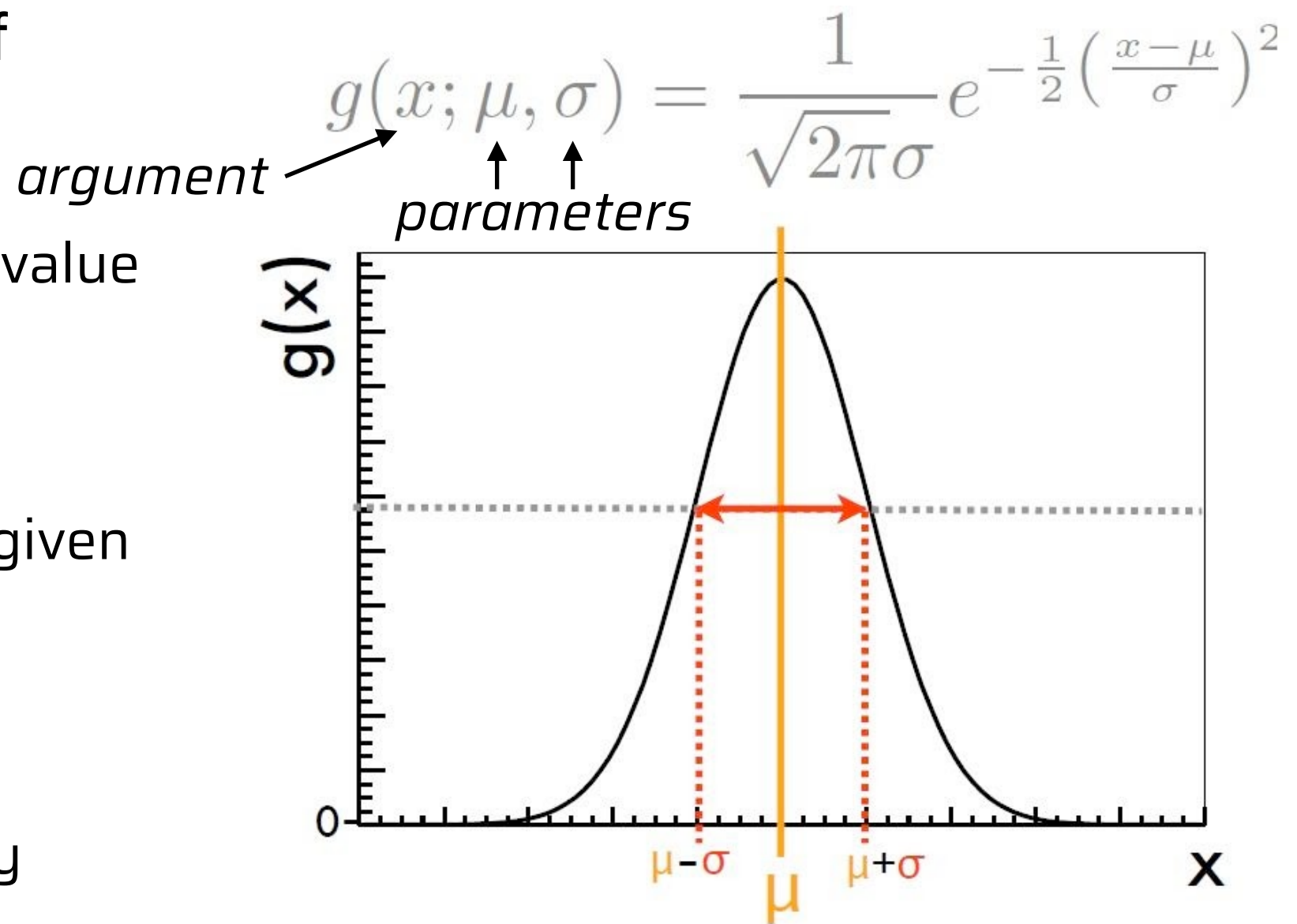
- Probabilities must add up to 1: integral on all possible values is unity (normalization)

$$\int_{-\infty}^{+\infty} p(x) dx = 1$$

Information encoded in a pdf

Example: Gaussian pdf

- Mode (peak), μ : most likely value
- Confidence region
 - Range encompassing a given probability (percentile)
 - Standard deviation σ : contains 68% probability



- Mean value, μ (median more representative if outliers/tails)

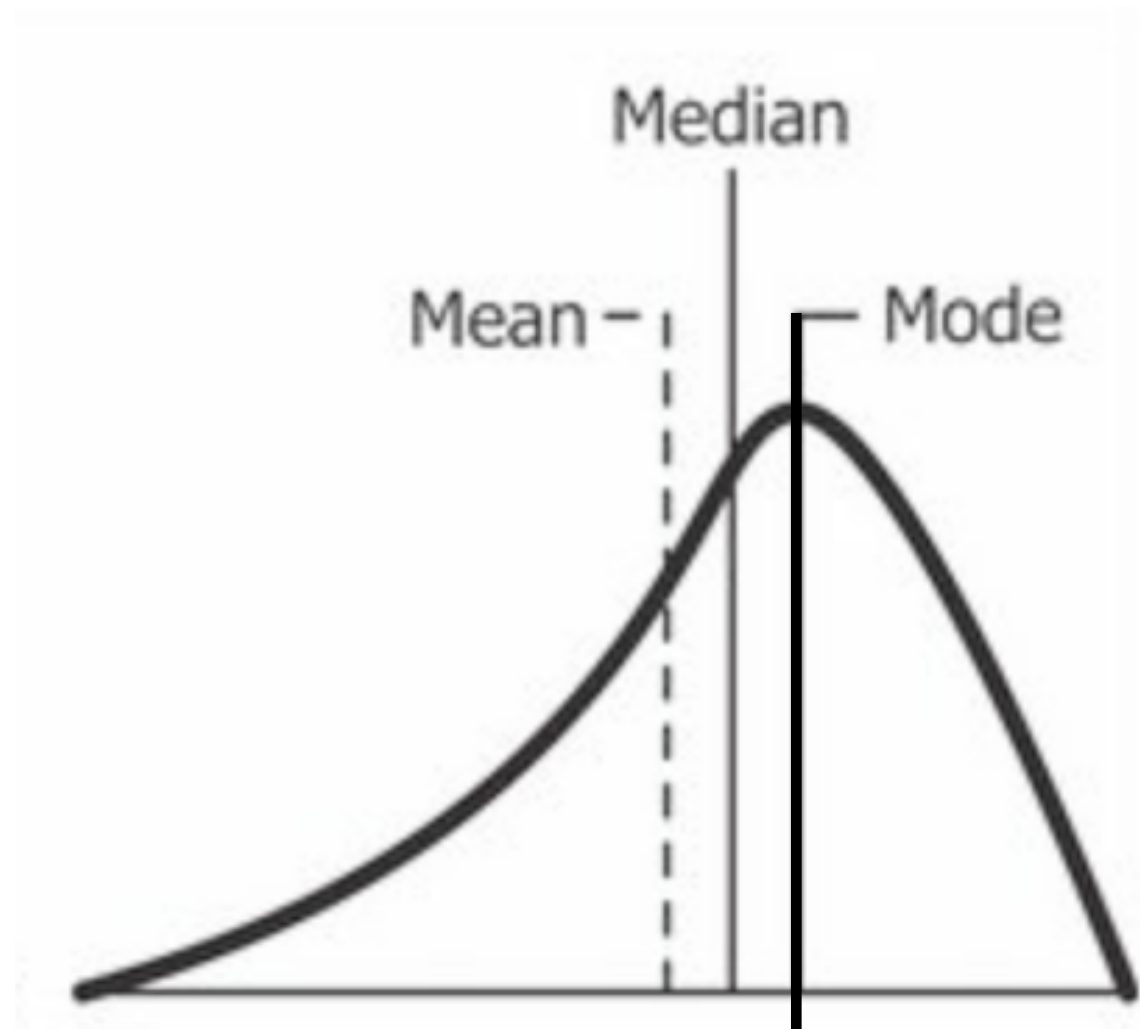
Whole shape of the pdf contains useful information (e.g. multiple peaks)

The pdf contains essentially all information we can get from a stochastic process

Information encoded in a pdf

Example: skewed pdf

- Mode (peak), μ : most likely value
- Confidence region
 - Range encompassing a given probability (percentile)
 - Standard deviation σ : contains 68% probability
- Mean value, μ (median more representative if outliers/tails)



Whole shape of the pdf contains useful information (e.g. multiple peaks)

The pdf contains essentially all information we can get from a stochastic process

Statistics, the Inverse Problem of Probability

For example, the exponential distribution.

$$f(x; \tau) = \frac{1}{\tau} e^{-x/\tau}$$

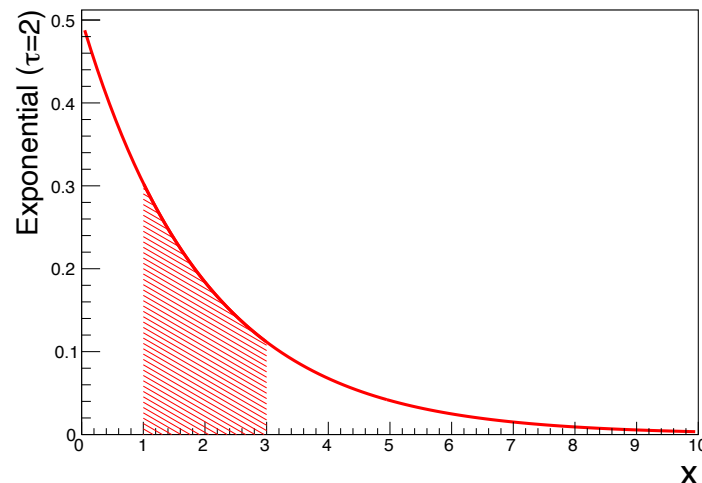
Probability: knowing τ , say something about x .

Statistics: knowing x , say something about τ .

Statistics, the Inverse Problem of Probability

Probability (easy):

- knowing that the half-life of a nucleus is $\tau = 2$ s
- what is the fraction of decays between 1 and 3 s?
- Unique solution: $P(1 < t < 3 | \tau = 2) = \int_1^3 f(x | \tau) dx = e^{-\frac{1}{2}} - e^{-\frac{3}{2}} = 0.3834$



Statistics, the Inverse Problem of Probability

Statistics (difficult):

- A random measurement of the exponential distribution gives: $x = 2.7$
- What can we say about the parameter τ ? $\tau = 2.7_{-0.8}^{+1.2}$ with 68% CL.
- The solution is “correct”, but **not unique**: many other correct solutions.
- Often there is no “good” solution: many **approximate** solutions.

The Likelihood Function

A parameter θ is to be estimated from measurements $\mathbf{x}$ according to some $f(x; \theta)$.

The probability to get x_1, x_2 after 2 independent measurements : $f(x_1; \theta) \times f(x_2; \theta)$

And the probability to obtain the n independent measurements $\mathbf{x} = \{x_i\}$ is

$$f(\mathbf{x} | \theta) = \prod_{i=1}^n f(x_i; \theta)$$

The Likelihood Function

A parameter θ is to be estimated from measurements $\mathbf{x}$ according to some $f(x; \theta)$.

The probability to get x_1, x_2 after 2 independent measurements : $f(x_1; \theta) \times f(x_2; \theta)$

And the probability to obtain the n independent measurements $\mathbf{x} = \{x_i\}$ is

$$f(\mathbf{x} | \theta) = \prod_{i=1}^n f(x_i; \theta)$$

This is the **likelihood function** but thought as function of θ , given the measurement $\{x_i\}$

$$\mathcal{L}(\theta) = \mathcal{L}(\theta | \mathbf{x}) \equiv \prod_{i=1}^n f(x_i; \theta)$$

The Likelihood Function

A parameter θ is to be estimated from measurements $\mathbf{x}$ according to some $f(x; \theta)$.

The probability to get x_1, x_2 after 2 independent measurements : $f(x_1; \theta) \times f(x_2; \theta)$

And the probability to obtain the n independent measurements $\mathbf{x} = \{x_i\}$ is

$$f(\mathbf{x} | \theta) = \prod_{i=1}^n f(x_i; \theta)$$

This is the **likelihood function** but thought as function of θ , given the measurement $\{x_i\}$

$$\mathcal{L}(\theta) = \mathcal{L}(\theta | \mathbf{x}) \equiv \prod_{i=1}^n f(x_i; \theta)$$

The notation $\mathcal{L}$ stresses that we mean fixed data $\{x_i\}$.

The Likelihood Function: the exponential distribution case

Consider n independent exponential measurements, $f(x; \tau) = \frac{1}{\tau} e^{-x/\tau}$

$$\mathcal{L}(\tau | \mathbf{x}) = \prod_{i=1}^n f(x_i | \tau) = \prod_{i=1}^n \frac{1}{\tau} e^{-x_i/\tau} = \frac{1}{\tau^n} e^{-\sum x_i/\tau}$$

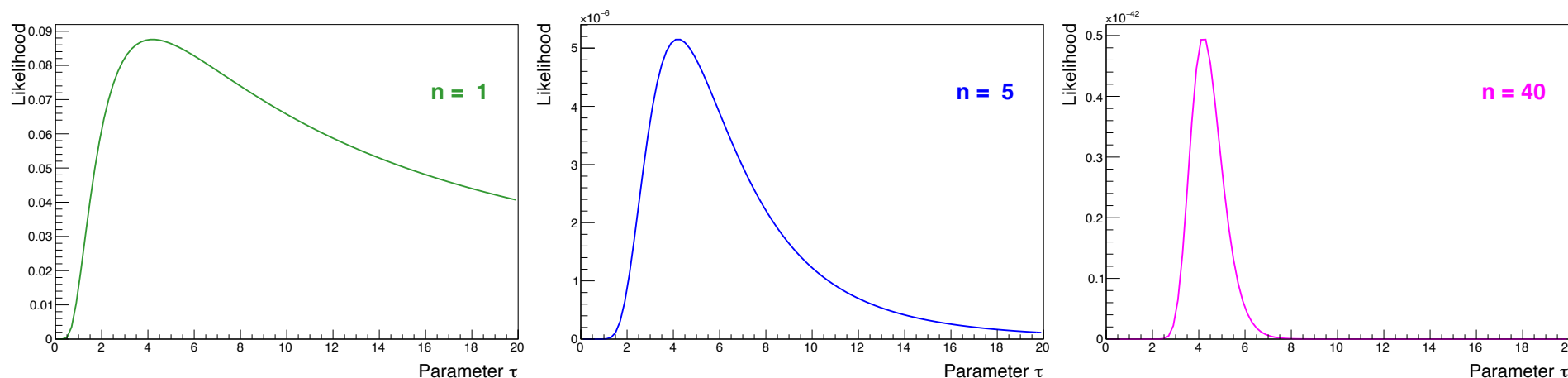
The likelihood expresses how likely the observed data $\mathbf{x}$ assuming a value of τ .

However $\mathcal{L}(\tau | \mathbf{x})$ is **NOT** a probability density on τ : $\int_0^\infty \mathcal{L}(\tau | \mathbf{x}) d\tau \neq 1$

It is only normalized in $\mathbf{x}$: $\int_0^\infty \mathcal{L}(\tau | \mathbf{x}) d\mathbf{x} = 1$

The Likelihood Function: the exponential distribution case

Plots $\mathcal{L}(\tau)$ for different values of n , at constant $\sum x_i/n$:



- Values of τ for which $\mathcal{L}$ is very small are ruled out.
- Best estimates of τ come from the region where $\mathcal{L}$ is largest.
- Precision of estimate for τ comes from width of $\mathcal{L}$ distribution.
- As n increases the admissible values of τ get restricted: precision increases.
- At large n the likelihood looks Gaussian.

The Maximum Likelihood Principle

The **Maximum Likelihood (ML)** estimate, $\hat{\theta}(\mathbf{x})$, is that which maximizes $\mathcal{L}$:

$$\mathcal{L}(\hat{\theta} | \mathbf{x}) \geq \mathcal{L}(\theta | \mathbf{x}) \quad \forall \theta$$

- The estimate $\hat{\theta}(\mathbf{x})$ is a statistic, a function of only the data, an observable.
- Numerical solution, maximizing $\mathcal{L}$ wrt θ at fixed $\mathbf{x}$: **unbinned fit**.
- For a few cases solution can be analytical, $\left. \frac{\partial \mathcal{L}(\theta | \mathbf{x})}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0$
- In general, $\log \mathcal{L}$ is easier to work with : $\frac{\partial \log \mathcal{L}}{\partial \theta} = 0 \Leftrightarrow \frac{\partial \mathcal{L}}{\partial \theta} = 0$

The Maximum Likelihood Principle: the exponential distribution case

The MLE for the exponential distribution, an example of analytical solution.

$$\mathcal{L}(\tau) = \frac{1}{\tau^n} e^{-\sum x_i / \tau}$$

$$\log \mathcal{L}(\tau) = -n \log \tau - \frac{\sum x_i}{\tau}$$

$$\frac{\partial \log \mathcal{L}(\tau)}{\partial \tau} = -\frac{n}{\tau} + \frac{\sum x_i}{\tau^2}$$

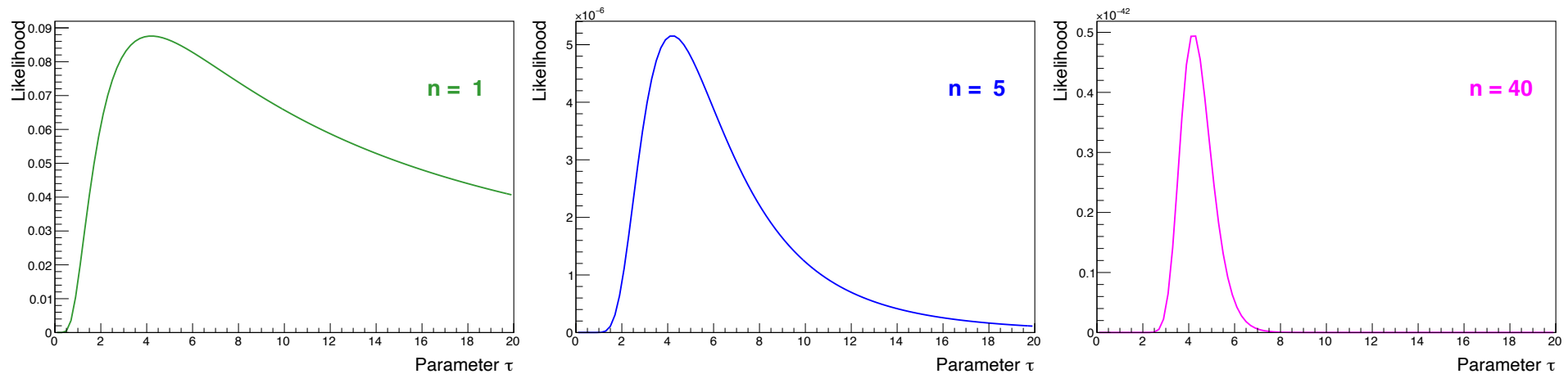
$$\left. \frac{\partial \log \mathcal{L}(\tau)}{\partial \tau} \right|_{\hat{\tau}} = 0 \Rightarrow -\frac{n}{\hat{\tau}} + \frac{\sum x_i}{\hat{\tau}^2} = 0$$

$$\hat{\tau} = \frac{\sum x_i}{n}$$

The Likelihood Function: extracting the Confidence Interval from $\mathcal{L}$

What about the uncertainty on the MLE parameters?

How to extract it from the width of the likelihood?



Confidence Intervals: the parabolic approximation

The MLE have the property of **asymptotic normality**: for n large, $\hat{\theta}$ tends to gaussian.

Exponential example: $\hat{\tau} = \frac{\sum x_i}{n}$, and $\sum x_i$ (sum of exponentials) tends to gaussian.

$$\mathcal{L}(\theta | \hat{\theta}) = f(\hat{\theta} | \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{\hat{\theta} - \theta}{\sigma} \right)^2}$$

$$\log \mathcal{L}(\theta | \hat{\theta}) = -\log \sqrt{2\pi}\sigma - \frac{1}{2} \left(\frac{\hat{\theta} - \theta}{\sigma} \right)^2$$

Repeating the experiment, $\hat{\theta}$ (random variable) fluctuates around θ (fixed-parameter).

Confidence Intervals: the parabolic approximation

The MLE have the property of **asymptotic normality**: for n large, $\hat{\theta}$ tends to gaussian.

Exponential example: $\hat{\tau} = \frac{\sum x_i}{n}$, and $\sum x_i$ (sum of exponentials) tends to gaussian.

$$\mathcal{L}(\theta | \hat{\theta}) = f(\hat{\theta} | \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\hat{\theta}-\theta}{\sigma}\right)^2}$$

$$\log \mathcal{L}(\theta | \hat{\theta}) = -\log \sqrt{2\pi}\sigma - \frac{1}{2}\left(\frac{\hat{\theta}-\theta}{\sigma}\right)^2$$

Repeating the experiment, $\hat{\theta}$ (random variable) fluctuates around θ (fixed-parameter).

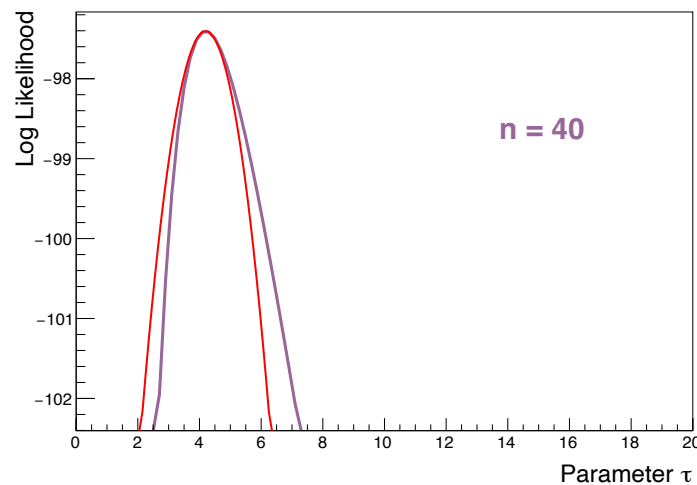
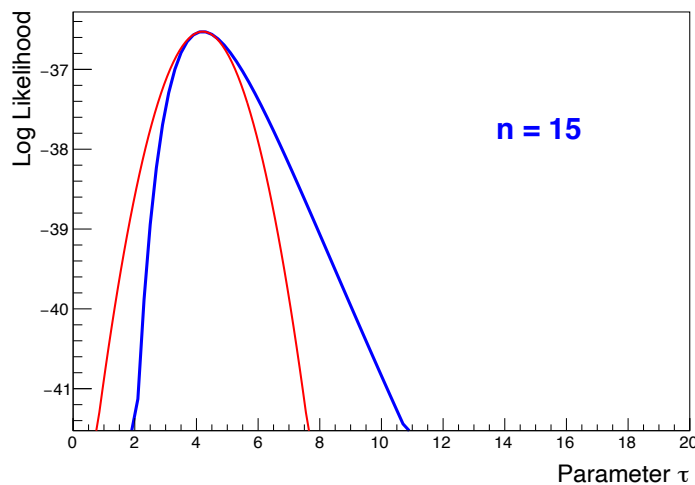
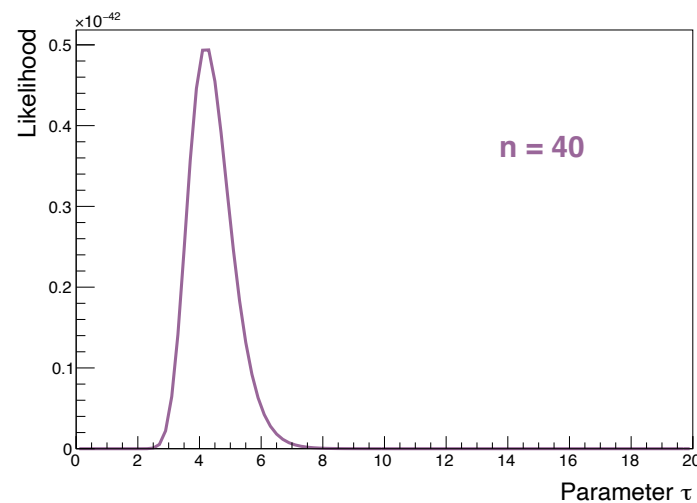
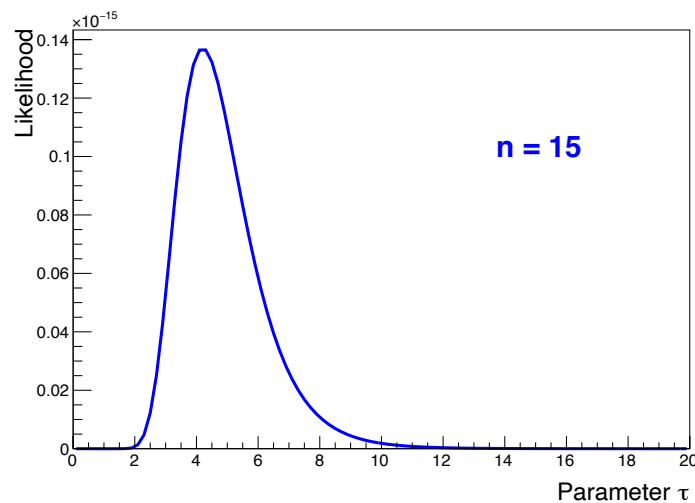
For **one** experiment:

$\mathcal{L}(\theta | \hat{\theta})$ is a gaussian centered on $\hat{\theta}$.

$\log \mathcal{L}(\theta | \hat{\theta})$ is a parabola centered on $\hat{\theta}$.

Confidence Intervals: the parabolic approximation

The gaussian (parabolic) approximation works better at higher n .



Confidence Intervals: the parabolic approximation

Within the parabolic approximation, the variance is extracted from the (usually numerically evaluated) second derivative of $\log \mathcal{L}$:

$$\mathcal{L}(\theta | \hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\sigma}\right)^2}$$

$$\log \mathcal{L}(\theta | \hat{\theta}) = -\log \sqrt{2\pi}\sigma - \frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\sigma}\right)^2$$

$$\frac{\partial \log \mathcal{L}(\theta | \hat{\theta})}{\partial \theta} = -\frac{\theta - \hat{\theta}}{\sigma^2}$$

$$\frac{\partial^2 \log \mathcal{L}(\theta | \hat{\theta})}{\partial \theta^2} = -\frac{1}{\sigma^2} \implies \sigma^2 = \left[-\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2} \right]_{\hat{\theta}}^{-1}$$

If several parameters:

$$\text{Cov}(\hat{\theta}_i, \hat{\theta}_j) = \left[-\frac{\partial^2 \log \mathcal{L}}{\partial \theta_i \partial \theta_j} \right]_{\hat{\theta}}^{-1}$$

Confidence Intervals: the parabolic approximation

- The parabolic approximation assumes validity of the asymptotic Gaussian regime.
- It is the default error calculation in most software packages.
- It only requires the second derivatives at the maximum, the hessian matrix.
- Accounts for correlations via the covariance matrix, the inverse of the hessian.

Likelihood Ratio Confidence Intervals

The Likelihood Function: Review

A parameter θ is to be estimated from measurements $\mathbf{x}$ according to some $f(x; \theta)$.

The probability to obtain the n independent measurements $\mathbf{x} = \{x_i\}$ is

$$f(\mathbf{x} | \theta) = \prod_{i=1}^n f(x_i; \theta)$$

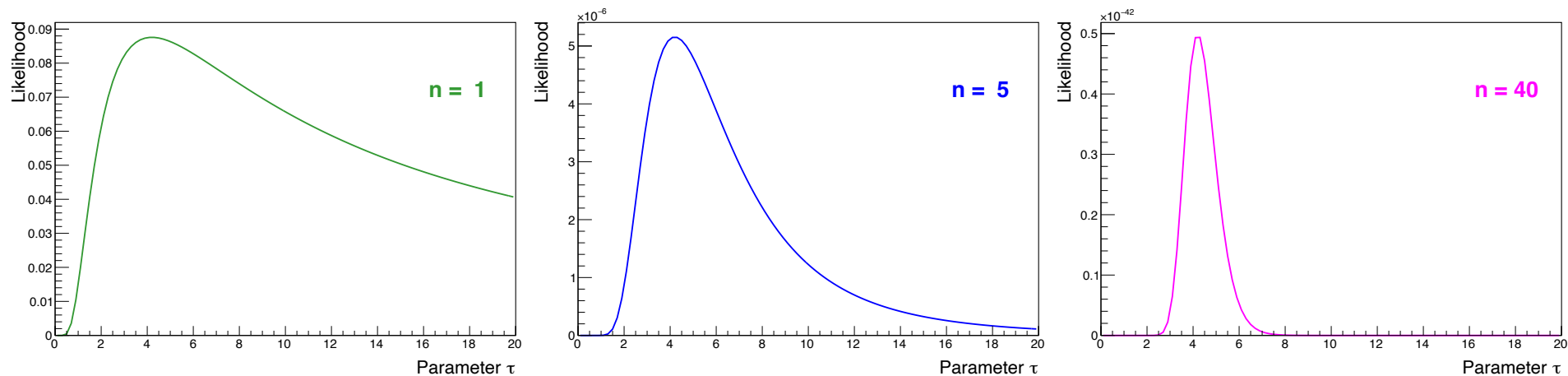
This is the **likelihood function** $\mathcal{L}(\theta | \mathbf{x})$, where $\mathcal{L}$ stresses we mean fixed data $\{x_i\}$.

$$\mathcal{L}(\theta) = \mathcal{L}(\theta | \mathbf{x}) \equiv \prod_{i=1}^n f(x_i; \theta)$$

The likelihood expresses how likely the observed data, $\mathbf{x}$, assuming a value of θ .

The Likelihood Function: Review

For the exponential case $f(x; \tau) = \frac{1}{\tau} e^{-x/\tau}$, the likelihood is $\mathcal{L}(\tau | \mathbf{x}) = \frac{1}{\tau^n} e^{-\sum x_i/\tau}$:



- The **Maximum Likelihood (ML)** estimate, $\hat{\theta}(\mathbf{x})$, is that which maximizes $\mathcal{L}$.
- For the case of the exponential, the MLE is: $\hat{\tau} = \frac{\sum x_i}{n}$
- The uncertainty in $\hat{\theta}(\mathbf{x})$ is related to the width of $\mathcal{L}$
- At large n the likelihood tends to a Gaussian (Central Limit Theorem).

Confidence Intervals: the Parabolic Approximation

Let's expand the likelihood $\log \mathcal{L}(\theta)$ in a series around $\hat{\theta}$:

$$\log \mathcal{L}(\theta) = \log \mathcal{L}(\hat{\theta}) + \left. \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} \right|_{\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2} \left. \frac{\partial^2 \log \mathcal{L}}{\partial \theta^2} \right|_{\hat{\theta}} (\theta - \hat{\theta})^2$$

In the linear term, the derivative is zero.

In the quadratic term, it is the inverse of the Cramér-Rao estimator of sigma:

$$\hat{\sigma}^2 = - \left. \frac{\partial^2 \log \mathcal{L}}{\partial \theta^2} \right|_{\hat{\theta}}$$

$$\log \mathcal{L}(\theta) = \log \mathcal{L}_{\max} - \frac{1}{2} \left(\frac{\theta - \hat{\theta}}{\hat{\sigma}} \right)^2$$

In the vicinity of the ML estimator, $\log \mathcal{L}$ is parabolic.

Confidence Intervals: the Parabolic Approximation

To see in which range the approximation is valid, the graphical interpretation of σ is useful.

$$\log \mathcal{L}_p(\theta) = \log \mathcal{L}_{\max} - \frac{1}{2} \left(\frac{\theta - \hat{\theta}}{\sigma} \right)^2$$

$$\log \mathcal{L}_p(\theta = \hat{\theta} \pm n\sigma) = \log \mathcal{L}_{\max} - \frac{n^2}{2}$$

$$\log \mathcal{L}_p(\hat{\theta} \pm n\sigma) = \log \mathcal{L}_{\max} - \frac{n^2}{2}$$

Confidence Intervals: the Parabolic Approximation

To see in which range the approximation is valid, the graphical interpretation of σ is useful.

$$\log \mathcal{L}_p(\theta) = \log \mathcal{L}_{\max} - \frac{1}{2} \left(\frac{\theta - \hat{\theta}}{\sigma} \right)^2$$

$$\log \mathcal{L}_p(\theta = \hat{\theta} \pm n\sigma) = \log \mathcal{L}_{\max} - \frac{n^2}{2}$$

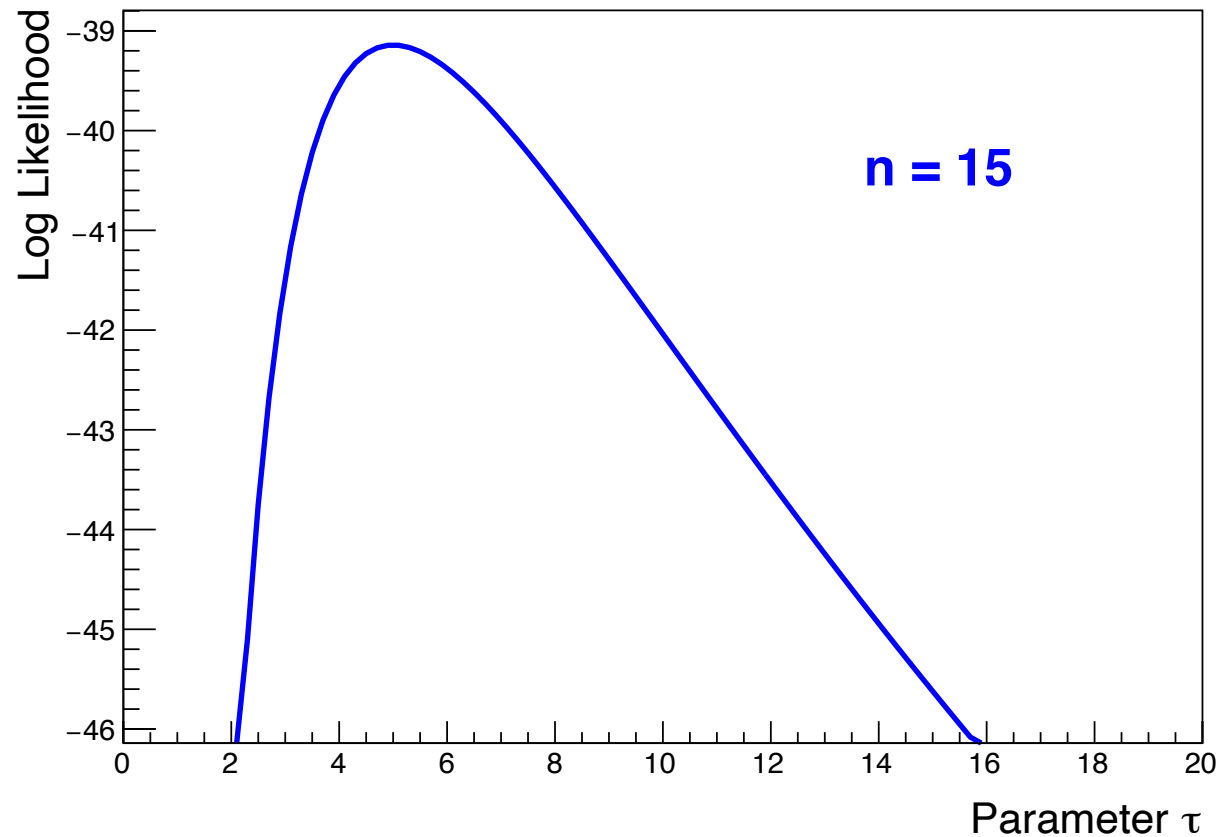
$$\log \mathcal{L}_p(\hat{\theta} \pm n\sigma) = \log \mathcal{L}_{\max} - \frac{n^2}{2}$$

$$68.3\% CL \ (\hat{\theta} \pm 1\sigma) : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - \frac{1^2}{2} \quad (0.5)$$

$$95.4\% CL \ (\hat{\theta} \pm 2\sigma) : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - \frac{2^2}{2} \quad (2.0)$$

$$99.7\% CL \ (\hat{\theta} \pm 3\sigma) : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - \frac{3^2}{2} \quad (4.5)$$

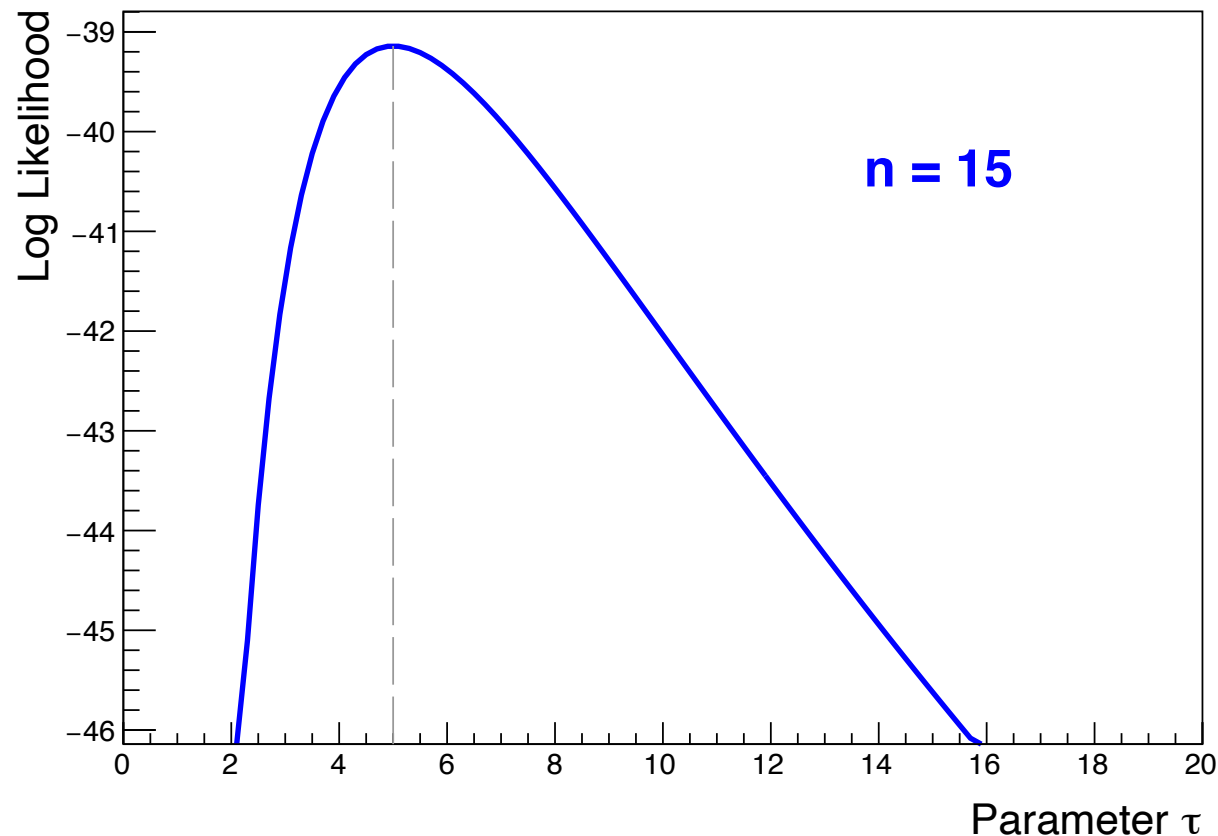
Confidence Intervals: the Parabolic Approximation



The log-likelihood for $n = 15$ measurements from an exponential distribution

$$\mathcal{L}(\tau | \mathbf{x}) = \frac{1}{\tau^n} e^{-\sum x_i / \tau}$$

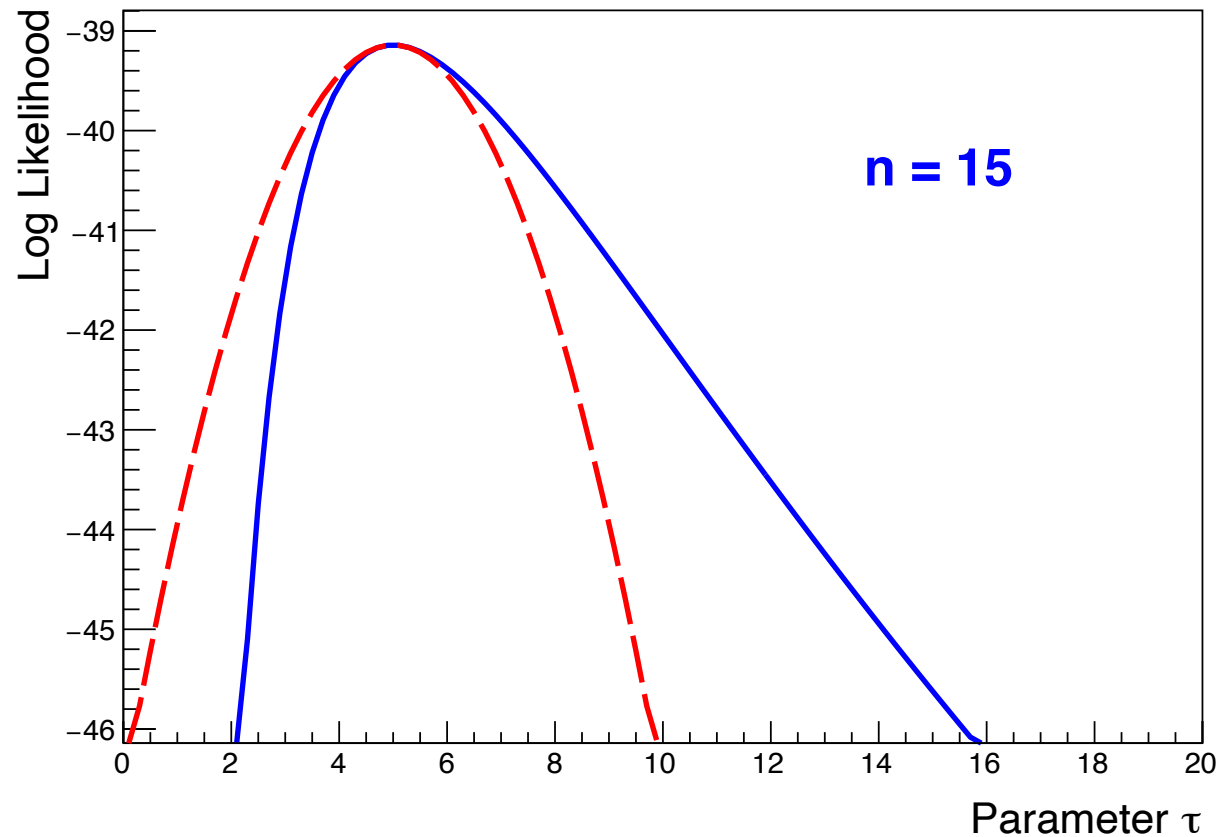
Confidence Intervals: the Parabolic Approximation



The log-likelihood for $n = 15$ measurements from an exponential distribution

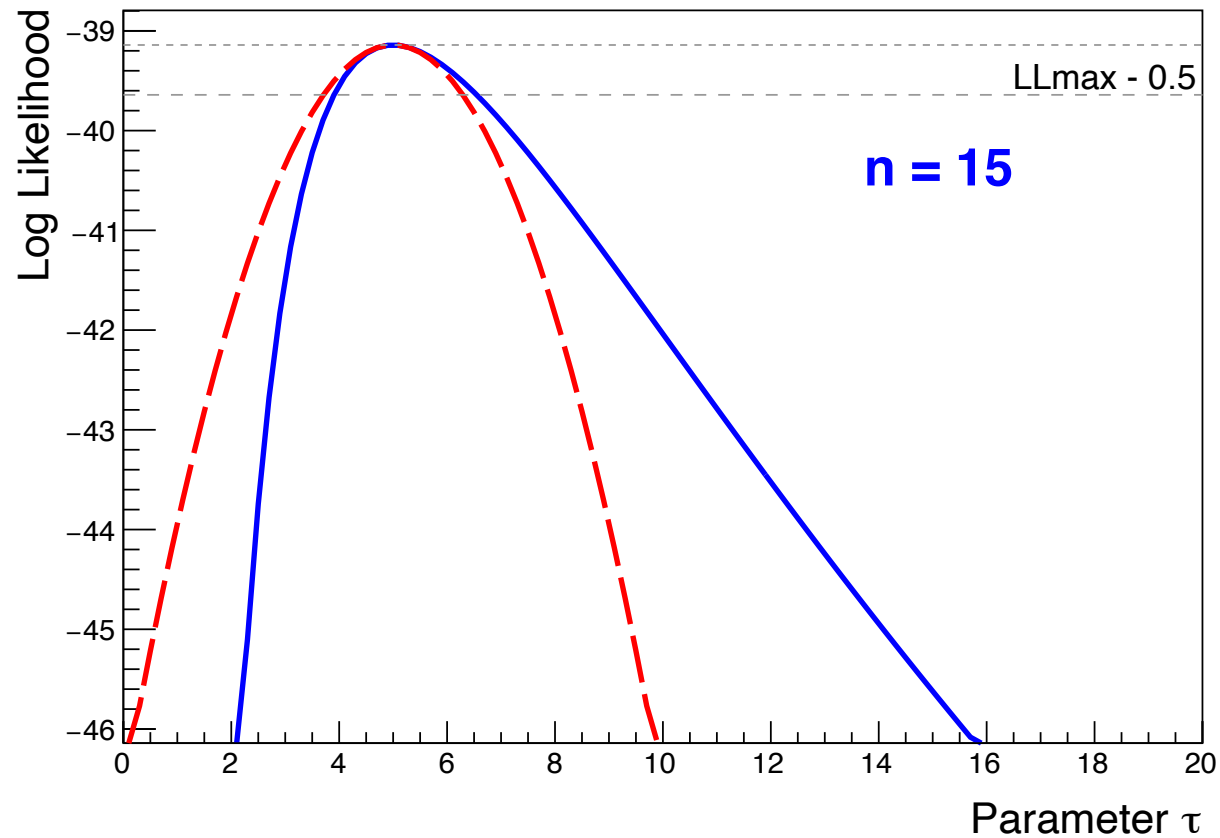
$$\mathcal{L}(\tau | \mathbf{x}) = \frac{1}{\tau^n} e^{-\sum x_i / \tau} \quad \text{The ML estimate in this example is } \hat{\tau} = \frac{\sum x_i}{n} = 5.0$$

Confidence Intervals: the Parabolic Approximation



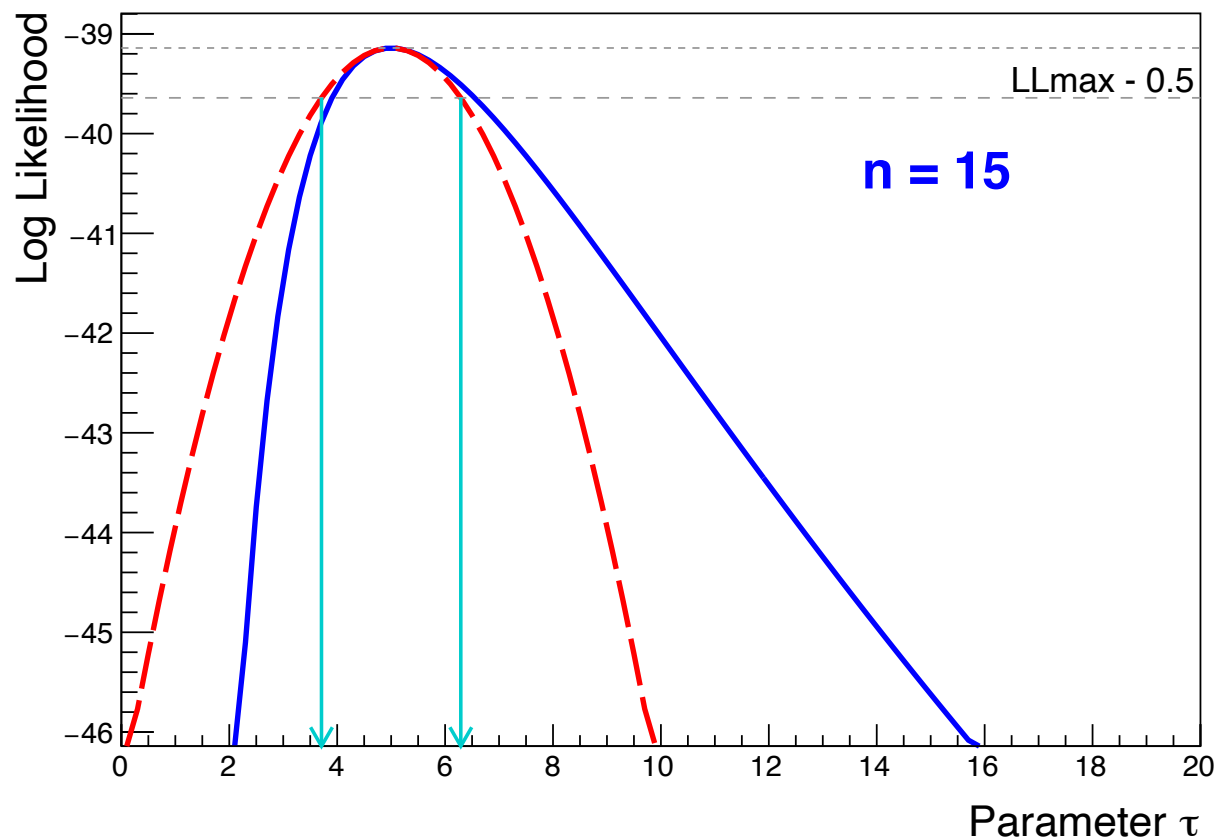
Trace the parabolic approximation $\mathcal{L}_p$ that matches $\mathcal{L}$ at the maximum ($\mathcal{L}_{\max}$)

Confidence Intervals: the Parabolic Approximation



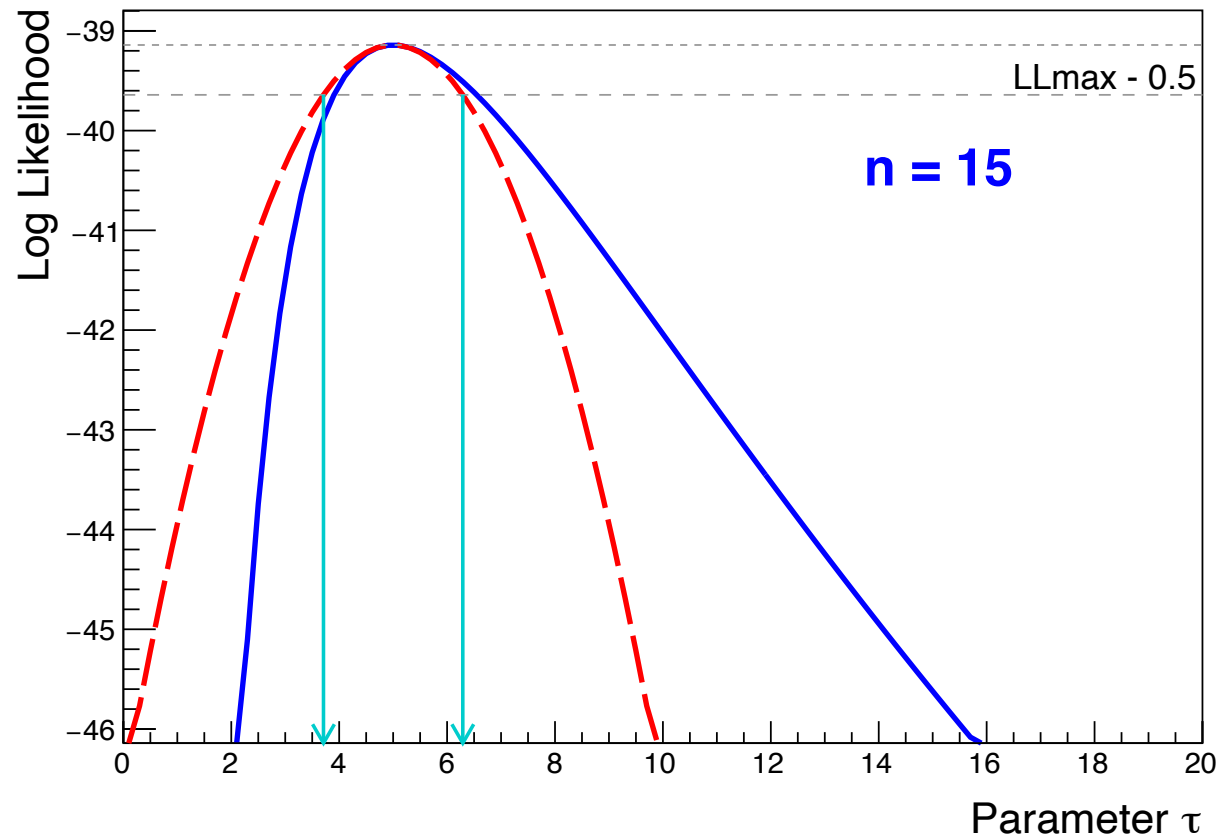
Draw a line at $\mathcal{L}_{\max} - 0.5$

Confidence Intervals: the Parabolic Approximation



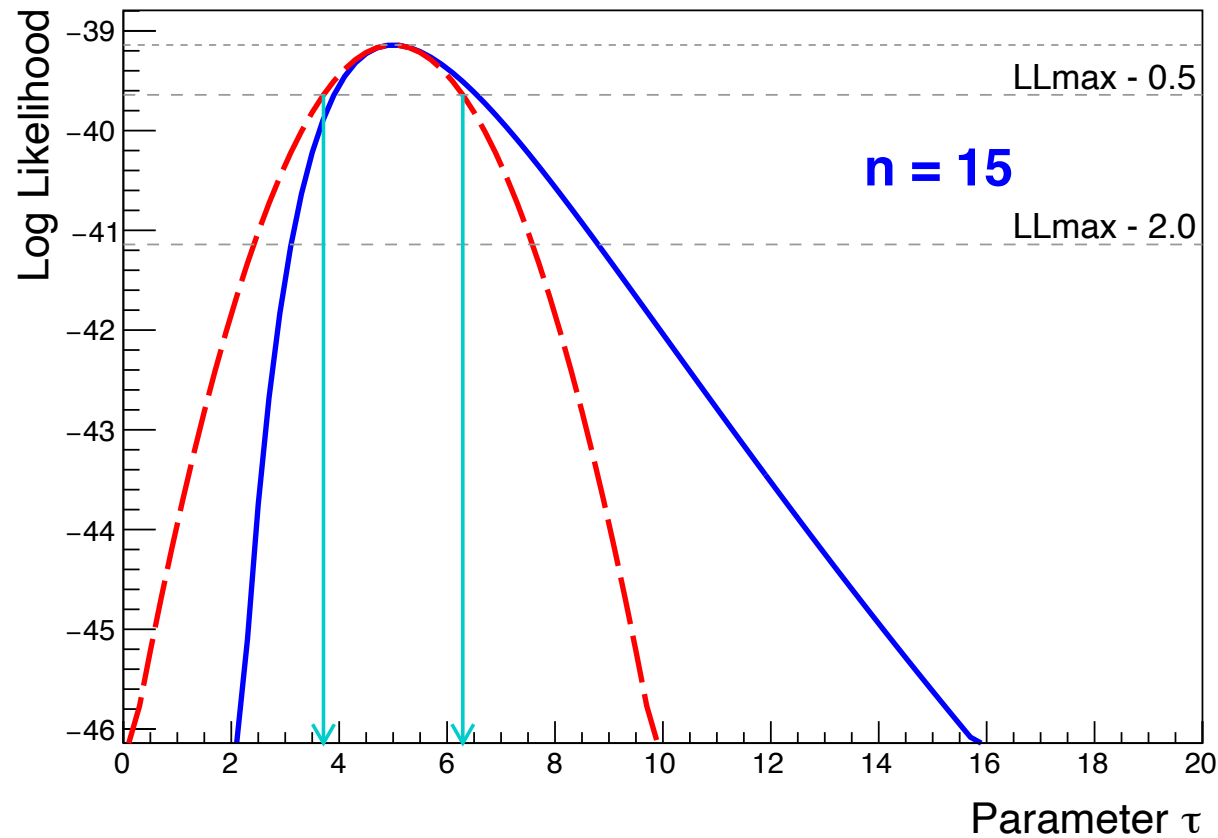
The 68.3% CL interval limits correspond to the points where $\mathcal{L}_p = \mathcal{L}_{\max} - 0.5$

Confidence Intervals: the Parabolic Approximation



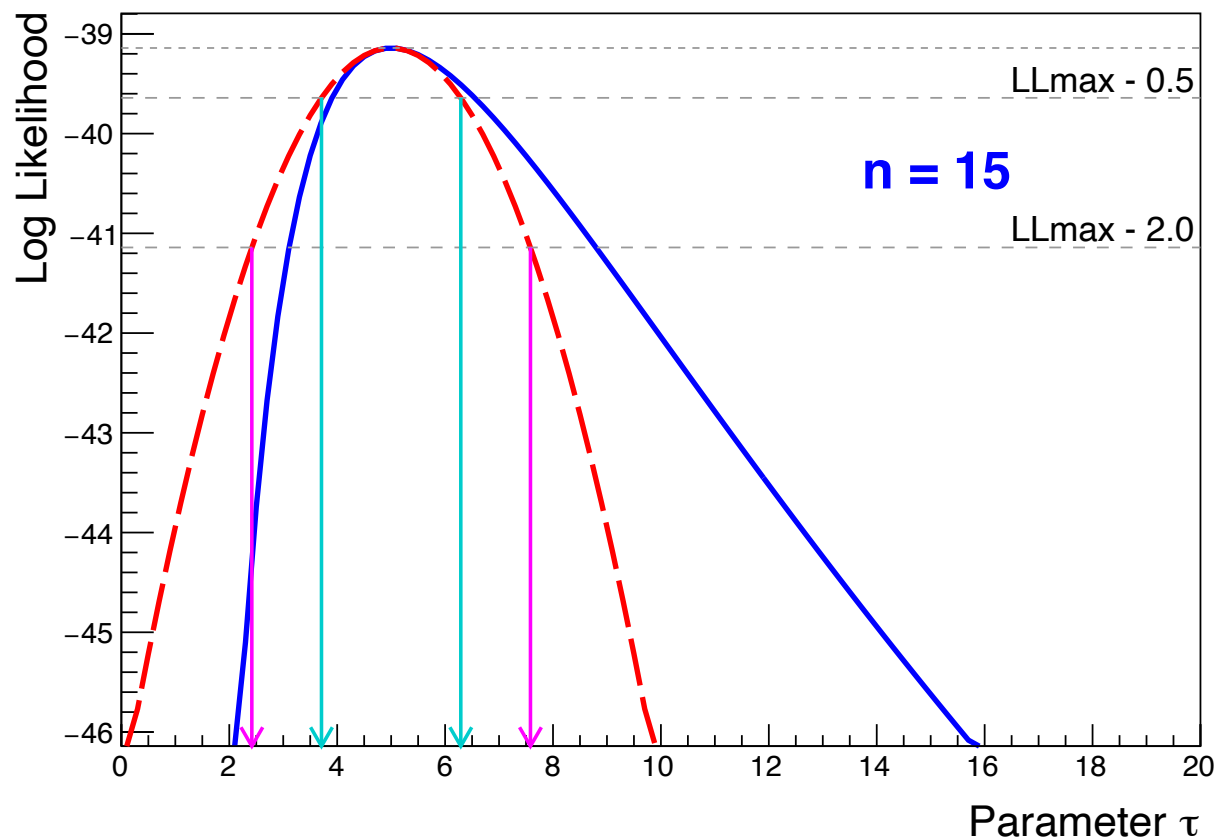
$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL \ (\hat{\tau} \pm 1\sigma) : \quad 5.0 \pm 1.3$$

Confidence Intervals: the Parabolic Approximation



$$\log \mathcal{L}_p(\tau_{\text{CI}}) = \log \mathcal{L}_{\text{max}} - 0.5 \quad 68.3\% \text{ CL } (\hat{\tau} \pm 1\sigma) : \quad 5.0 \pm 1.3$$

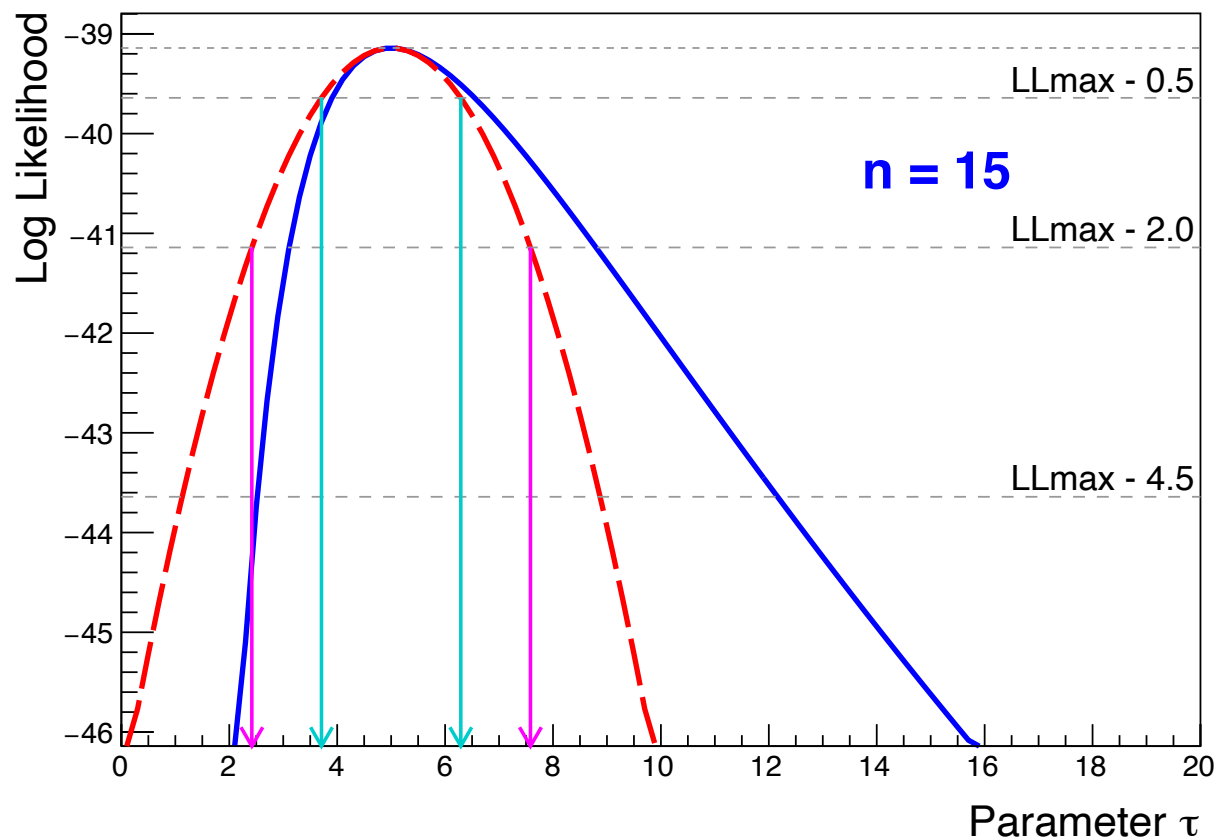
Confidence Intervals: the Parabolic Approximation



$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL \ (\hat{\tau} \pm 1\sigma) : \quad 5.0 \pm 1.3$$

$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 2.0 \quad 95.4\% CL \ (\hat{\tau} \pm 2\sigma) : \quad 5.0 \pm 2.6$$

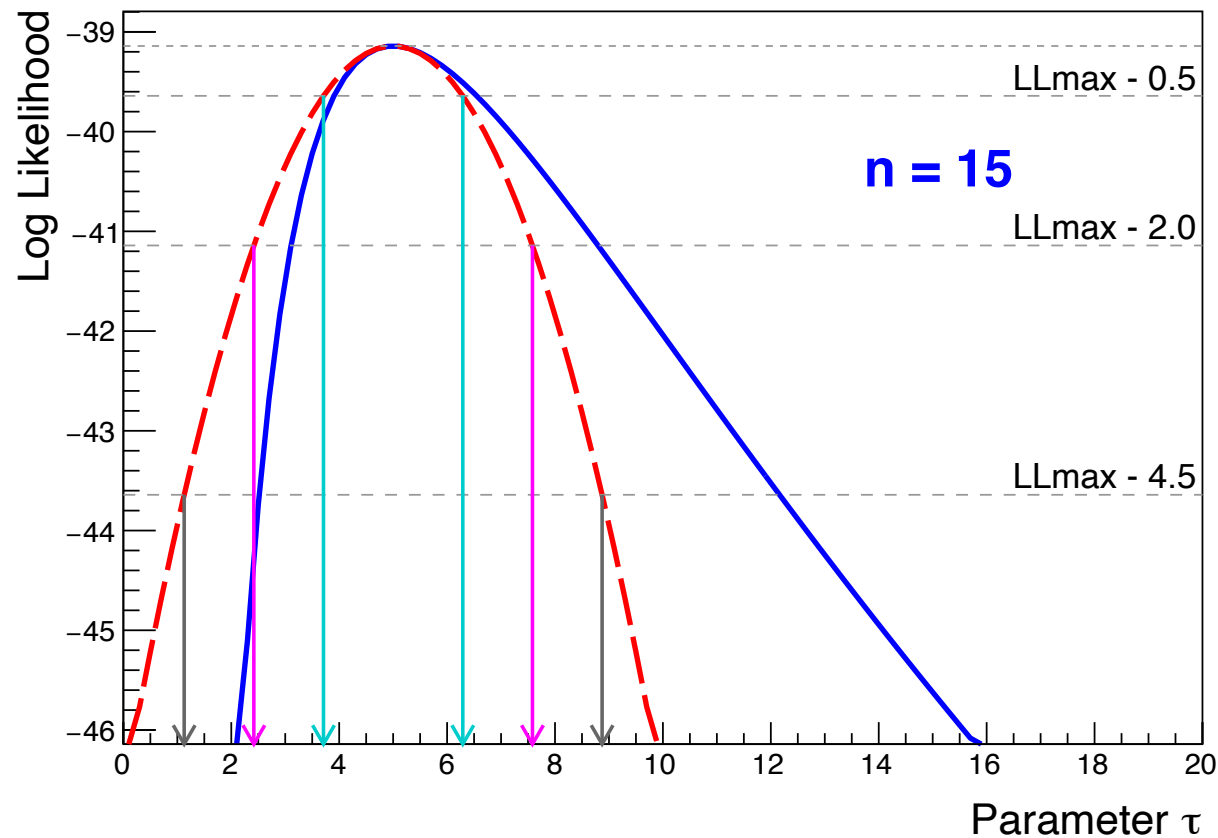
Confidence Intervals: the Parabolic Approximation



$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL \ (\hat{\tau} \pm 1\sigma) : \quad 5.0 \pm 1.3$$

$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 2.0 \quad 95.4\% CL \ (\hat{\tau} \pm 2\sigma) : \quad 5.0 \pm 2.6$$

Confidence Intervals: the Parabolic Approximation

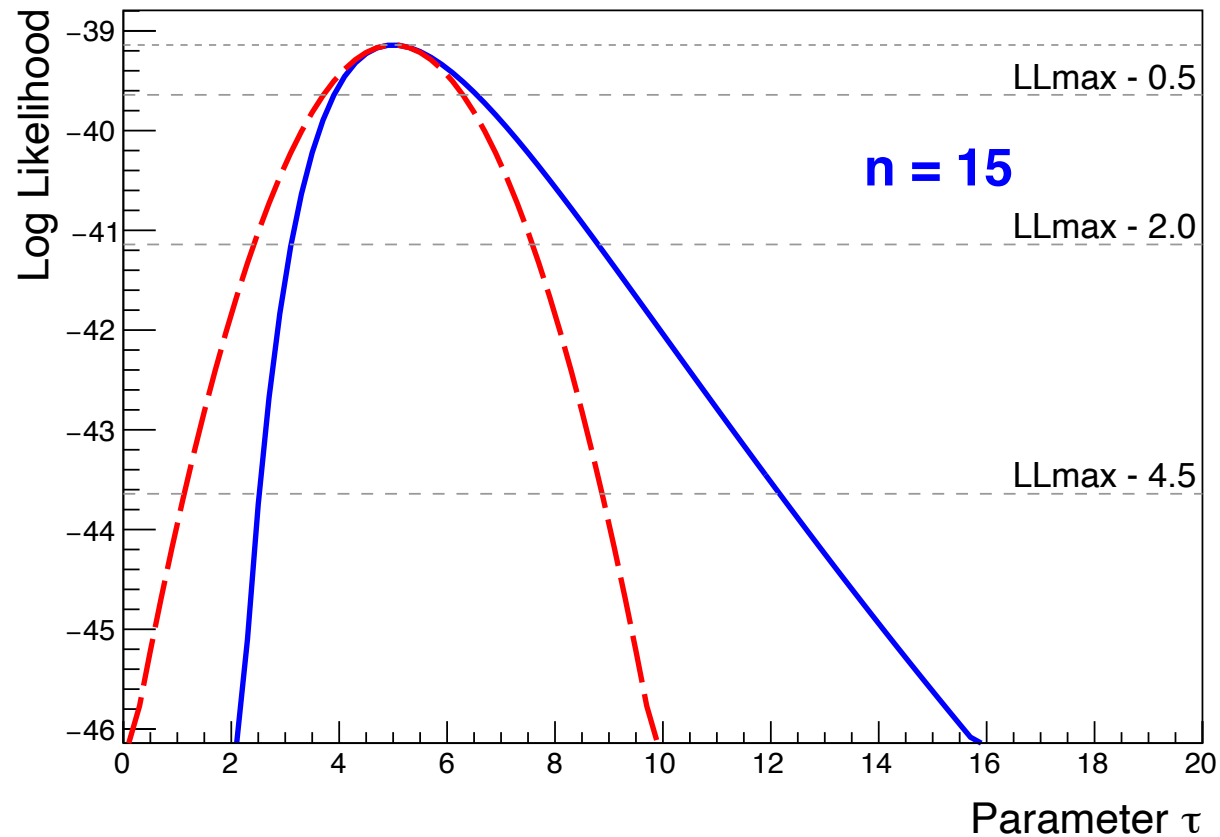


$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL \quad (\hat{\tau} \pm 1\sigma) : \quad 5.0 \pm 1.3$$

$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 2.0 \quad 95.4\% CL \quad (\hat{\tau} \pm 2\sigma) : \quad 5.0 \pm 2.6$$

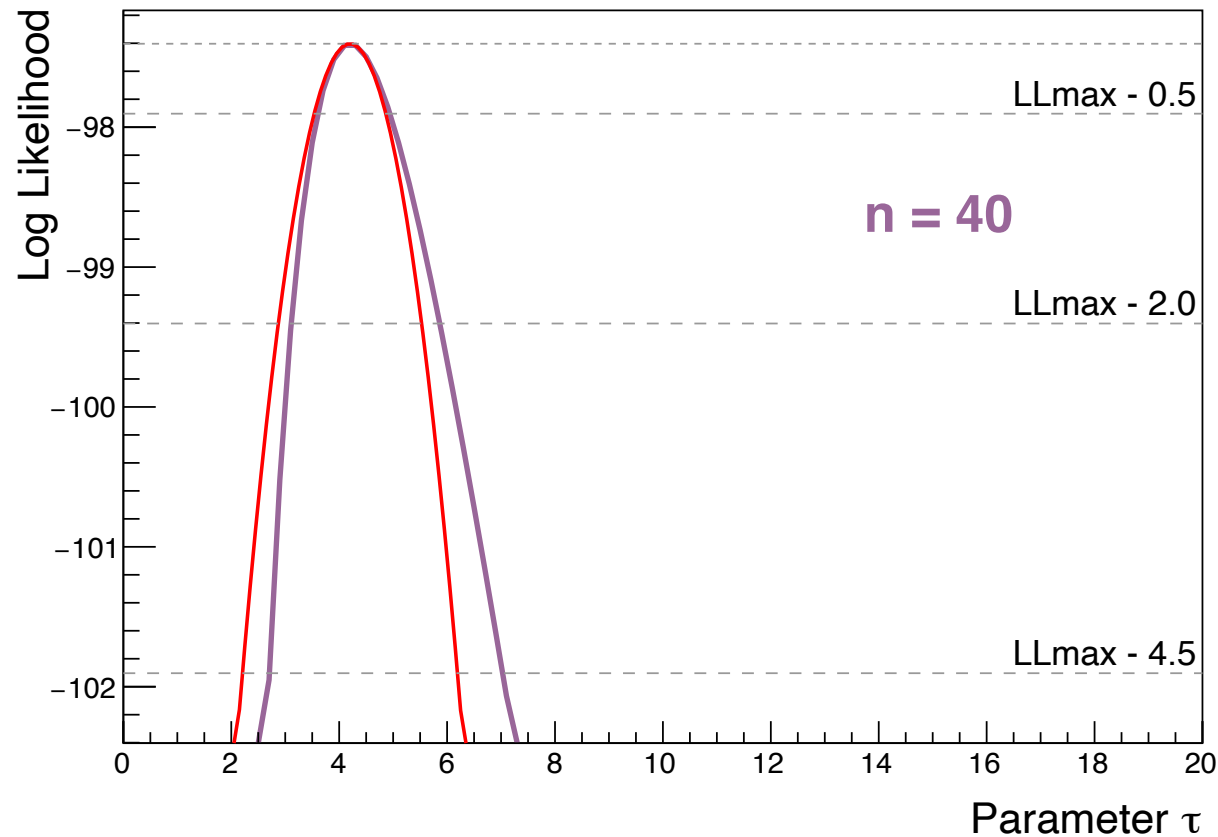
$$\log \mathcal{L}_p(\tau_{CI}) = \log \mathcal{L}_{\max} - 4.5 \quad 99.7\% CL \quad (\hat{\tau} \pm 3\sigma) : \quad 5.0 \pm 3.9$$

Confidence Intervals: the Parabolic Approximation



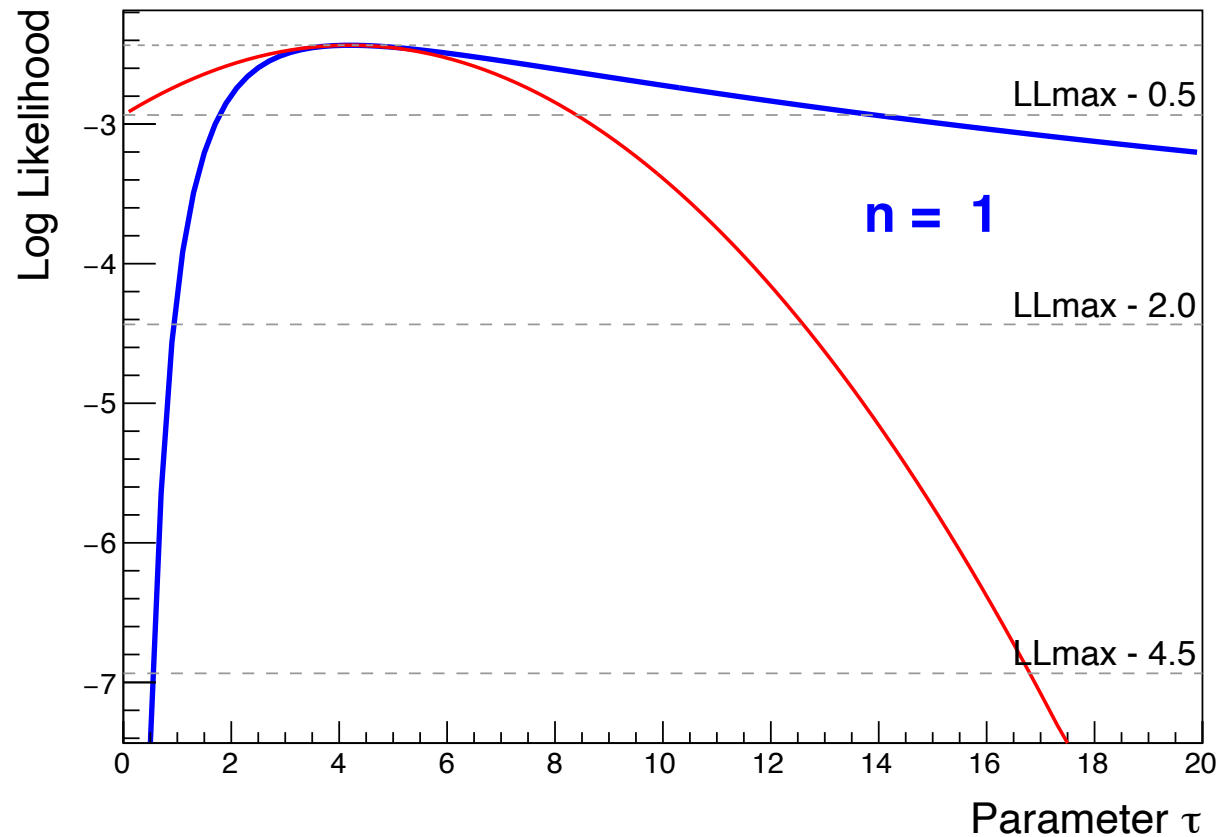
For $n = 15$, the parabolic approximation works fine up to $\pm 1\sigma$

Confidence Intervals: the Parabolic Approximation



For $n = 40$, the parabolic approximation should work fine for $\pm 1\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$

Intervalos de Confianza: la aproximación parabólica



For $n = 1$, the parabolic approximation does not work even for $\pm 1\sigma$

Confidence Intervals: the Parabolic Approximation

The recipe for the LLR intervals (also called MINOS confidence intervals) is very simple.

Take the recipe for the parabolic approximation:

$$68.3\% CL : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - 0.5$$

$$95.4\% CL : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - 2.0$$

$$99.7\% CL : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - 4.5$$

Confidence Intervals: the Parabolic Approximation

The recipe for the LLR intervals (also called MINOS confidence intervals) is very simple.

Take the recipe for the parabolic approximation:

$$68.3\% CL : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - 0.5$$

$$95.4\% CL : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - 2.0$$

$$99.7\% CL : \log \mathcal{L}_p(\theta_{CI}) = \log \mathcal{L}_{\max} - 4.5$$

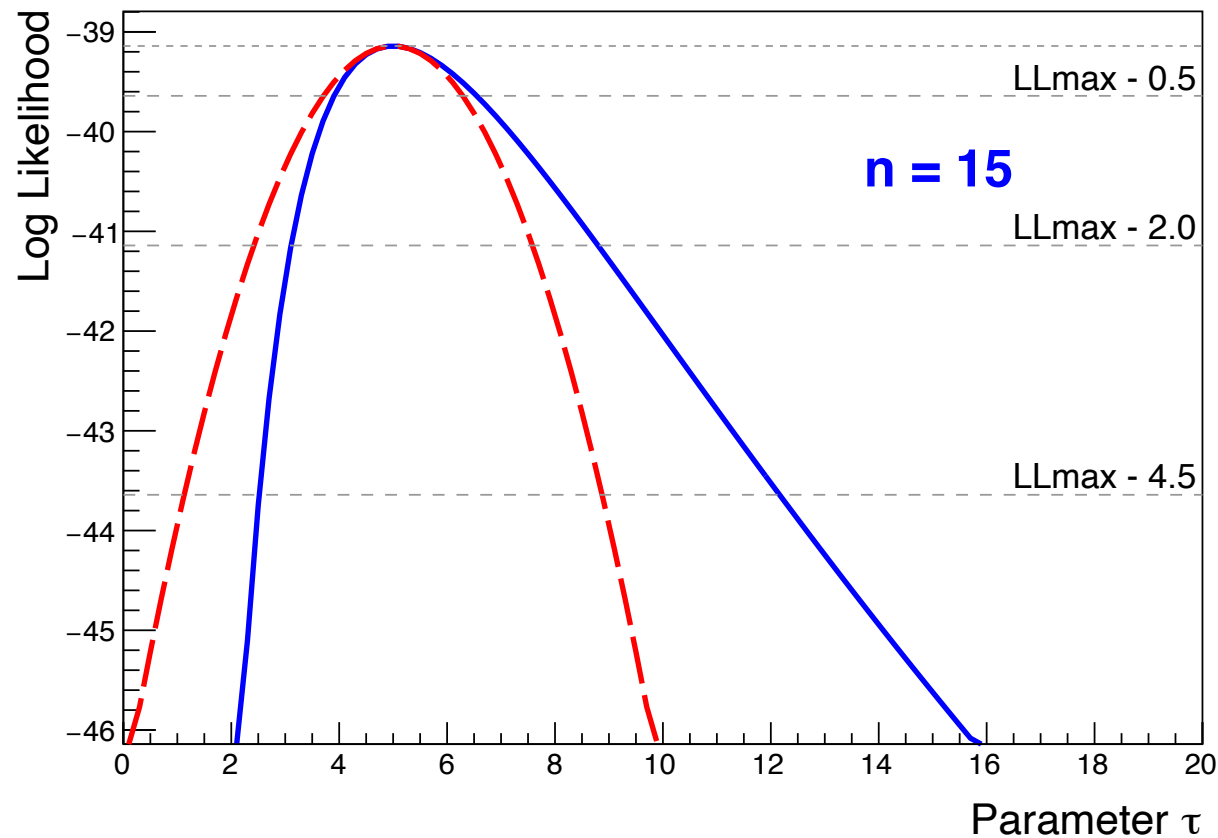
and use the correct likelihood instead of its parabolic approximation:

$$68.3\% CL : \log \mathcal{L}(\theta_{CI}) = \log \mathcal{L}_{\max} - 0.5$$

$$95.4\% CL : \log \mathcal{L}(\theta_{CI}) = \log \mathcal{L}_{\max} - 2.0$$

$$99.7\% CL : \log \mathcal{L}(\theta_{CI}) = \log \mathcal{L}_{\max} - 4.5$$

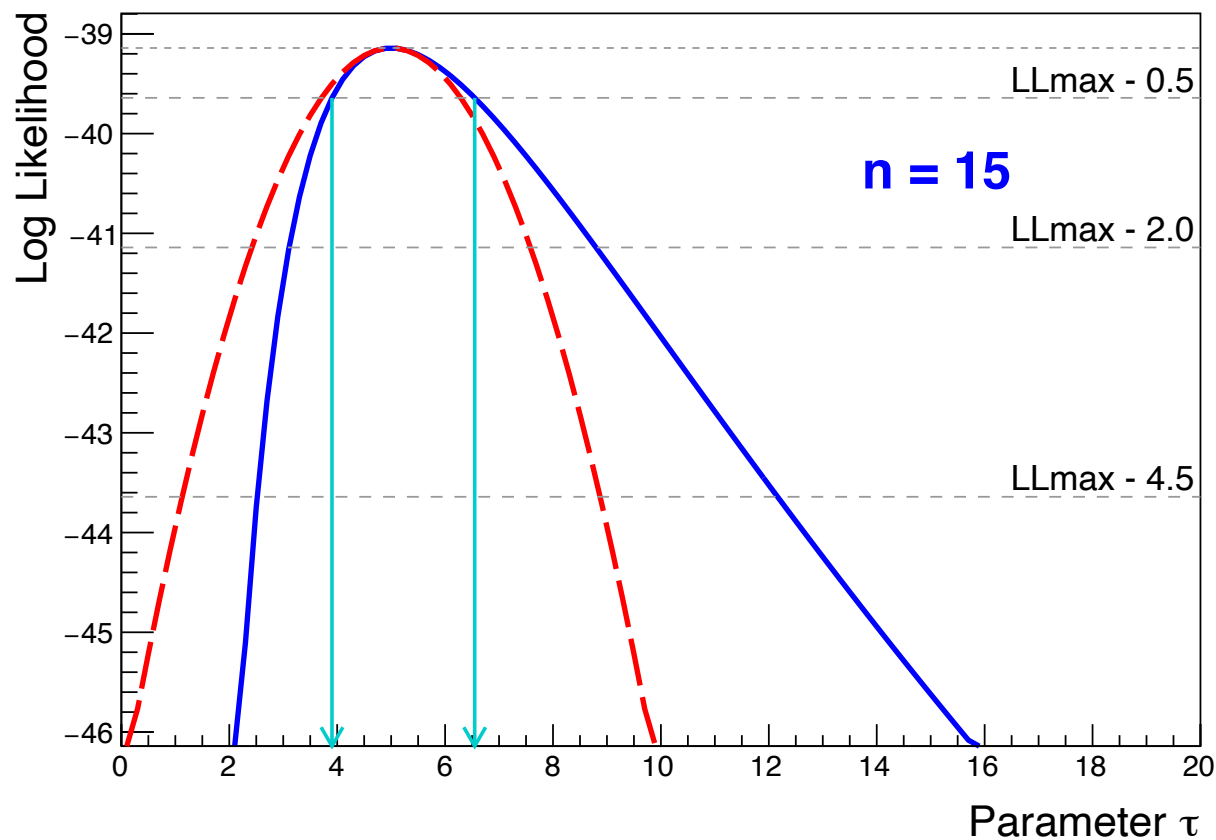
Confidence Intervals: the log-likelihood ratio (LLR) approximation



Start with the full likelihood $\mathcal{L}(\tau)$.

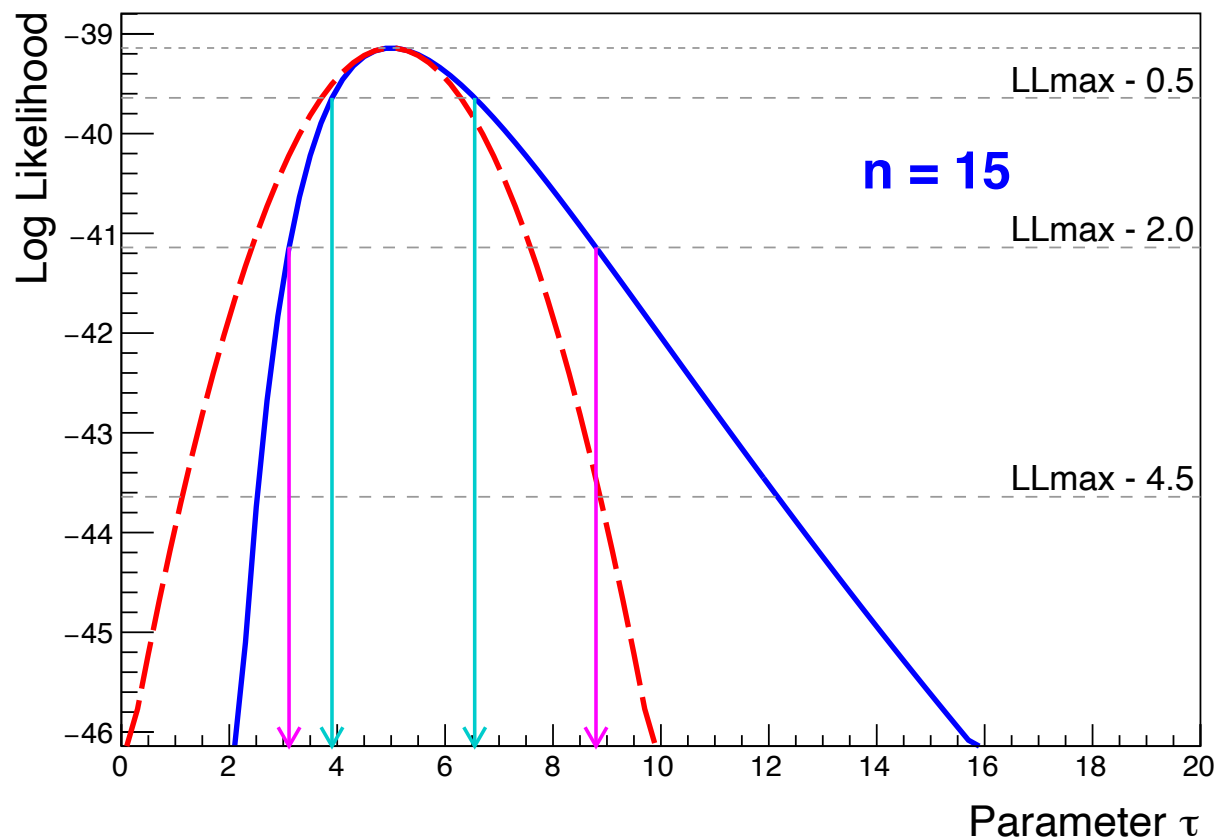
Draw lines at $\mathcal{L}_{\max} - \frac{1^2}{2}$, $\mathcal{L}_{\max} - \frac{2^2}{2}$, and $\mathcal{L}_{\max} - \frac{3^2}{2}$.

Confidence Intervals: the log-likelihood ratio (LLR) approximation



$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL : \quad 5.0^{+1.5}_{-1.1} \quad (5.0 \pm 1.3)$$

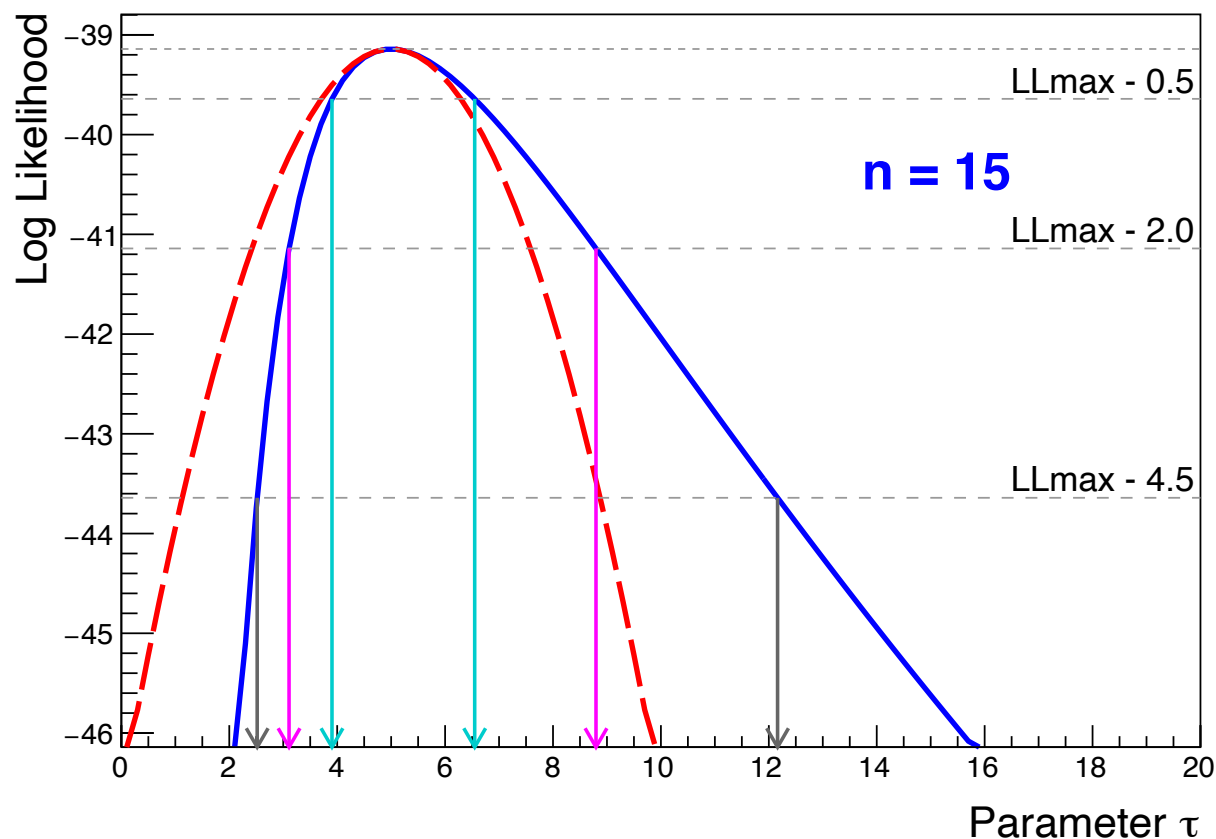
Confidence Intervals: the log-likelihood ratio (LLR) approximation



$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL : \quad 5.0^{+1.5}_{-1.1} \quad (5.0 \pm 1.3)$$

$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 2.0 \quad 95.4\% CL : \quad 5.0^{+3.8}_{-1.9} \quad (5.0 \pm 2.6)$$

Confidence Intervals: the log-likelihood ratio (LLR) approximation

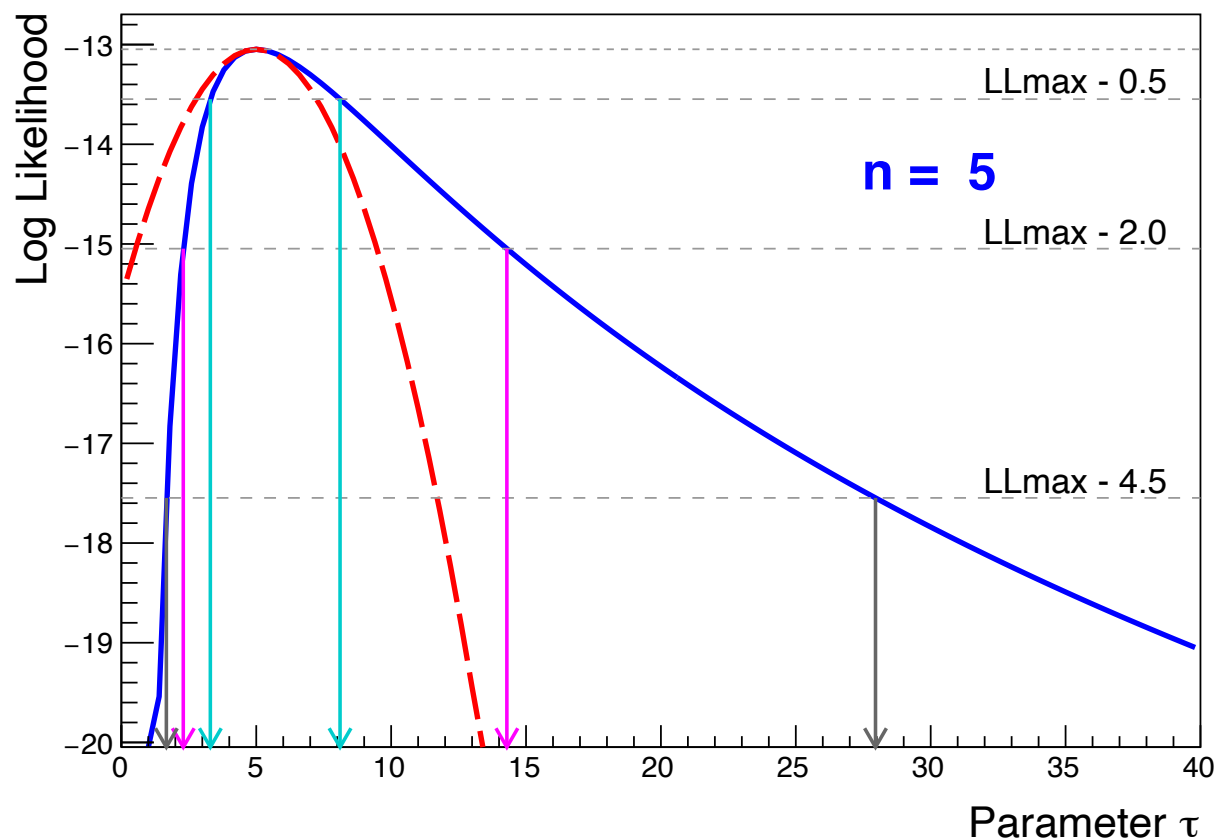


$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL : \quad 5.0^{+1.5}_{-1.1} \quad (5.0 \pm 1.3)$$

$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 2.0 \quad 95.4\% CL : \quad 5.0^{+3.8}_{-1.9} \quad (5.0 \pm 2.6)$$

$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 4.5 \quad 99.7\% CL : \quad 5.0^{+7.2}_{-2.5} \quad (5.0 \pm 3.9)$$

Confidence Intervals: the log-likelihood ratio (LLR) approximation



$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 0.5 \quad 68.3\% CL : \quad 5.0^{+3.1}_{-1.7} \quad (5.0 \pm 2.2)$$

$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 2.0 \quad 95.4\% CL : \quad 5.0^{+9.3}_{-2.7} \quad (5.0 \pm 4.5)$$

$$\log \mathcal{L}(\tau_{CI}) = \log \mathcal{L}_{\max} - 4.5 \quad 99.7\% CL : \quad 5.0^{+23.3}_{-3.5} \quad (5.0 \pm 6.7)$$

Confidence Intervals: the log-likelihood ratio (LLR) approximation

Assessing the quality of the likelihood intervals:

- Generate a sample of n measurements from the exponential $\tau = 2$ distribution.
- Calculate for this sample the parabolic or LLR interval at a given CL.
- Check if the true value ($\tau = 2$) is included in (covered by) the interval.
- Repeat the experiment many times ($N = 10^6$, for instance).
- Measure coverage: fraction of experiments where CI contains the true value.
- Verify if the coverage of the interval is equal to its Confidence Level.

The Likelihood Function: extracting the Confidence Interval from $\mathcal{L}$

Coverage results for n measurements from an exponential distribution with $\tau = 2$.

The CI are obtained at each CL using the parabolic and the LLR approximations.

	68.3%		95.4%		99.7%	
n	Parab.	LLR	Parab.	LLR	Parab.	LLR
1	60.5	64.4	71.8	80.7	77.1	81.2
2	66.6	66.3	79.8	92.6	85.9	93.6
3	67.7	66.9	83.4	95.1	89.6	97.4
4	68.4	67.2	85.4	94.6	92.1	99.1
5	68.1	68.0	87.0	94.9	93.1	99.5
6	68.2	68.5	87.8	95.2	94.7	99.6
8			90.1	95.3	95.4	99.6
10			90.5	95.0	96.3	99.6
15			92.3	95.5	97.0	99.6
20			92.8	95.2	97.8	99.6
30			93.7	95.4	98.3	99.7
40			93.8	95.2	98.8	99.7

Maximum Likelihood Principle → Least Squares Method

For n measurements from a distribution $f(x | \theta)$, the likelihood is $\mathcal{L}(\theta) = \prod f(x_i | \theta)$.

When $f(x | \theta)$ is Gaussian, the ML principle yields the method of least squares,

also called χ^2 minimization:

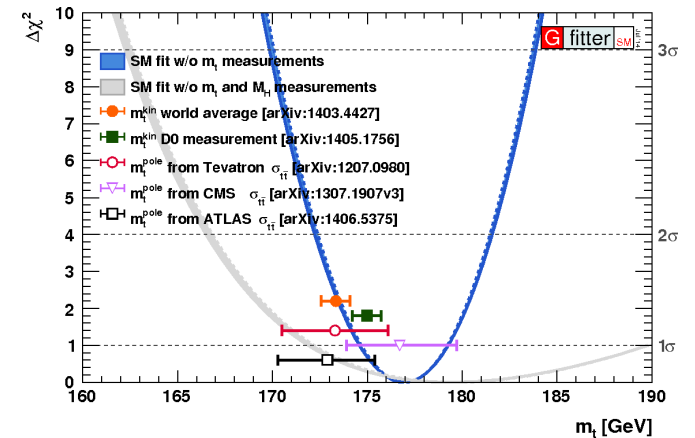
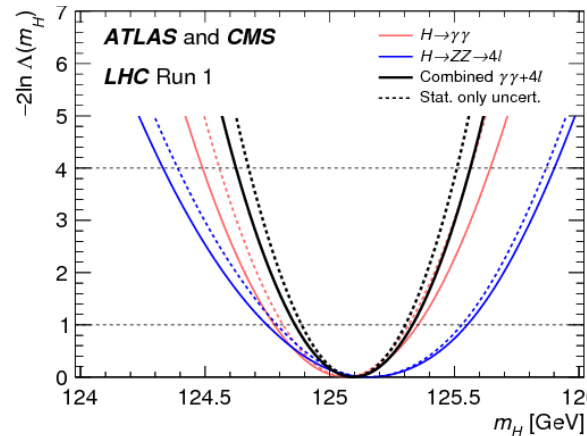
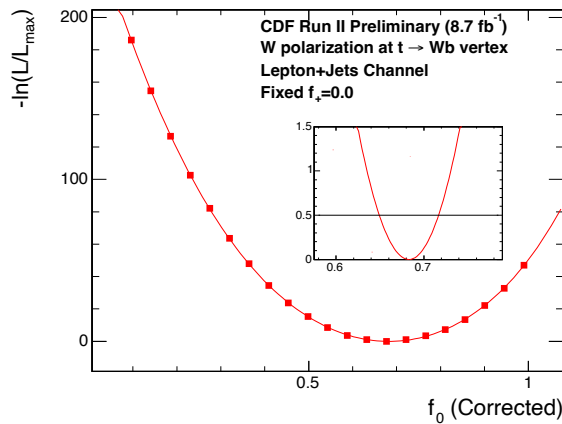
$$\mathcal{L}(\theta) = \prod_{i=1}^n f(x_i | \theta) = C \prod_{i=1}^n e^{-\frac{1}{2} \left(\frac{x_i - \mu(\theta)}{\sigma_i} \right)^2} \implies -2 \log \mathcal{L}(\theta) = \sum_{i=1}^n \left(\frac{x_i - \mu(\theta)}{\sigma_i} \right)^2 + C'$$

Maximizing $\mathcal{L}(\theta)$ is the same as minimizing the sum of n gaussians squared,

which is the $\chi^2_{(n)}$ distribution.

Alternative nomenclatures used in the literature

$\log \mathcal{L}$	$-\log \mathcal{L}$	$-2 \log \mathcal{L}$	χ^2
$\log \mathcal{L}_{\max} - \frac{1}{2}$	$(-\log \mathcal{L}_{\max}) + \frac{1}{2}$	$(-2 \log \mathcal{L}_{\max}) + 1$	$\chi^2_{\min} + 1$
$\log \mathcal{L}_{\max} - \frac{4}{2}$	$(-\log \mathcal{L}_{\max}) + \frac{4}{2}$	$(-2 \log \mathcal{L}_{\max}) + 4$	$\chi^2_{\min} + 4$
$\log \mathcal{L}_{\max} - \frac{9}{2}$	$(-\log \mathcal{L}_{\max}) + \frac{9}{2}$	$(-2 \log \mathcal{L}_{\max}) + 9$	$\chi^2_{\min} + 9$



$$-\ln(\mathcal{L}/\mathcal{L}_{\max}) = 0.5 \Rightarrow -\ln \mathcal{L} + \ln \mathcal{L}_{\max} = 0.5 \Rightarrow \ln \mathcal{L} = \ln \mathcal{L}_{\max} - \frac{1}{2}$$

$$\Lambda = \mathcal{L}/\mathcal{L}_{\max} \Rightarrow -2 \ln \Lambda = 1 \Rightarrow -2 \ln \mathcal{L} + 2 \ln \mathcal{L}_{\max} = 1 \Rightarrow -2 \ln \mathcal{L} = (-2 \ln \mathcal{L}_{\max}) + 1$$

LLR Confidence Intervals Theory: One–dimensional case

The 68.3% LLR Confidence Interval is composed of the values of θ that satisfy :

$$\log \mathcal{L}(\theta) \geq \log \mathcal{L}(\hat{\theta}) - \frac{1}{2}$$

This is a consequence of *Wilks' Theorem*:

“Twice the negative log-likelihood ratio is a random variable with distribution $\chi^2_{(1)}$ ”

$$q(\hat{\theta}) \equiv -2 \log \frac{\mathcal{L}(\theta | \mathbf{x})}{\mathcal{L}(\hat{\theta} | \mathbf{x})} \sim \chi^2_{(1)}$$

Notice that the likelihood ratio $\frac{\mathcal{L}(\theta | \mathbf{x})}{\mathcal{L}(\hat{\theta} | \mathbf{x})}$ takes values in the $(0, 1]$ range.

LLR Confidence Intervals Theory: One-dimensional case

Wilk's can also be written as

$$-2 \log \frac{\mathcal{L}(\theta | \mathbf{x})}{\mathcal{L}(\hat{\theta} | \mathbf{x})} = 2 \log \mathcal{L}(\hat{\theta}) - 2 \log \mathcal{L}(\theta) = q(\hat{\theta}) \sim \chi^2_{(1)}$$

How far can get $\hat{\theta}$ from θ with 68% probability?

$$P \left[\left(2 \log \mathcal{L}(\hat{\theta}) - 2 \log \mathcal{L}(\theta) \right) \leq c \right] = 0.68$$

LLR Confidence Intervals Theory: One-dimensional case

Wilk's can also be written as

$$-2 \log \frac{\mathcal{L}(\theta | \mathbf{x})}{\mathcal{L}(\hat{\theta} | \mathbf{x})} = 2 \log \mathcal{L}(\hat{\theta}) - 2 \log \mathcal{L}(\theta) = q(\hat{\theta}) \sim \chi^2_{(1)}$$

How far can get $\hat{\theta}$ from θ with 68% probability?

$$P \left[\left(2 \log \mathcal{L}(\hat{\theta}) - 2 \log \mathcal{L}(\theta) \right) \leq c \right] = 0.68$$

Using Wilks' Theorem:

$$P \left[q(\hat{\theta}) \leq c \right] = 0.68 \quad \Rightarrow \quad c = 1$$

LLR Confidence Intervals Theory: One-dimensional case

Wilk's can also be written as

$$-2 \log \frac{\mathcal{L}(\theta | \mathbf{x})}{\mathcal{L}(\hat{\theta} | \mathbf{x})} = 2 \log \mathcal{L}(\hat{\theta}) - 2 \log \mathcal{L}(\theta) = q(\hat{\theta}) \sim \chi^2_{(1)}$$

How far can get $\hat{\theta}$ from θ with 68% probability?

$$P \left[\left(2 \log \mathcal{L}(\hat{\theta}) - 2 \log \mathcal{L}(\theta) \right) \leq c \right] = 0.68$$

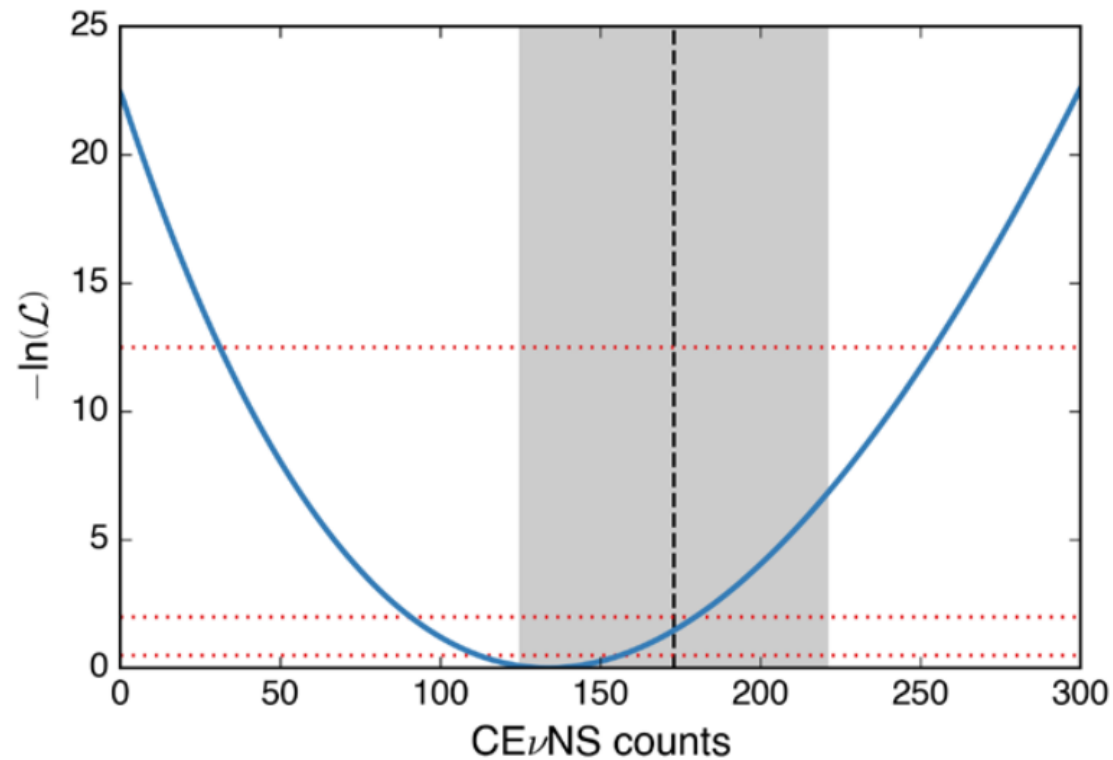
Using Wilks' Theorem:

$$P \left[q(\hat{\theta}) \leq c \right] = 0.68 \quad \Rightarrow \quad c = 1$$

$$P \left[\left(2 \log \mathcal{L}(\hat{\theta}) - 2 \log \mathcal{L}(\theta) \right) \leq 1 \right] = 0.68$$

$$68\% CL : \quad \log \mathcal{L}(\theta) \geq \log \mathcal{L}(\hat{\theta}) - \frac{1}{2}$$

LLR intervals: example



Negative profile log-likelihood for the number of CE ν NS events present in CsI[Na] data. Likelihood values are shifted to 0.

This result is within the 68% confidence band of the SM prediction shown as a shaded region and a vertical dashed line. The 68%, 95%, and 99.9999% confidence levels (1 sigma, 2 sigma, and 5 sigma) of the fit are shown as ascending horizontal dotted lines.

Two-dimensional LLR intervals

Two parameters: $L(\boldsymbol{\theta} | \boldsymbol{x}) = \mathcal{L}(\theta_1, \theta_2 | \boldsymbol{x})$

The ML estimate is similar:

$$\mathcal{L}(\hat{\theta}_1, \hat{\theta}_2 | \boldsymbol{x}) \geq \mathcal{L}(\theta_1, \theta_2 | \boldsymbol{x}) \quad \forall \theta_1, \theta_2$$

The intervals are now 2-dim contours:

$$\log \mathcal{L}(\theta_1, \theta_2) \geq \log \mathcal{L}(\hat{\theta}_1, \hat{\theta}_2) - \frac{1}{2} \chi_{2,CL}^2$$

$$\text{where } \chi_{2,CL}^2 = \begin{cases} 2.30 & 68.3\% CL \\ 6.16 & 95.4\% CL \\ 11.62 & 99.7\% CL \end{cases}$$

PHYSICAL REVIEW D **87**, 111101(R) (2013)

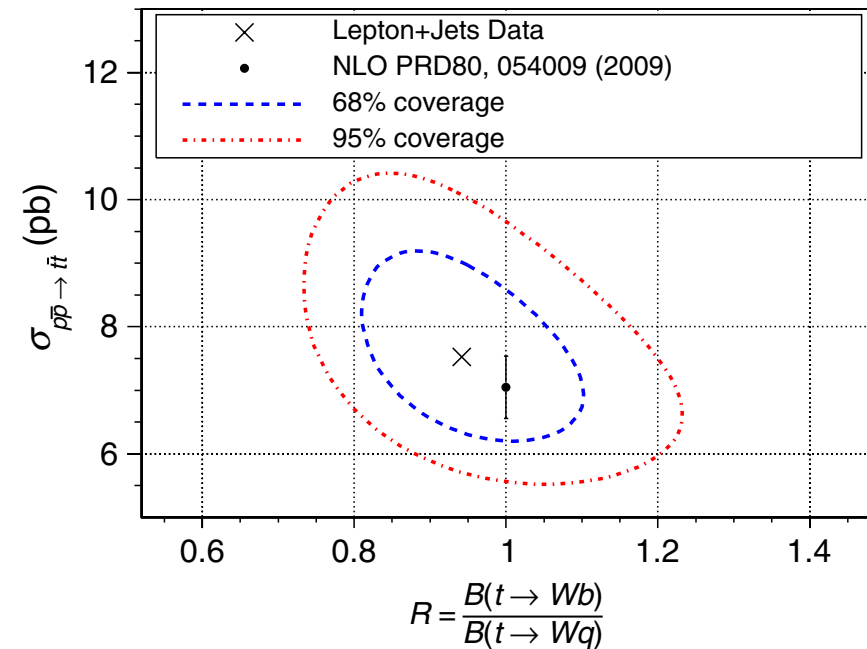


FIG. 2 (color online). The fit results for the simultaneous measurement of R and $\sigma_{t\bar{t}}$. The X-cross corresponds to the maximum of the likelihood; the point with error bar to the NLO cross section calculation. The two-dimensional confidence regions are shown as well.

LLR Confidence Intervals Theory: Multi-dimensional case

Wilk's for several parameters, $\boldsymbol{\theta} = \{\theta_k\}$

$$-2 \log \frac{\mathcal{L}(\boldsymbol{\theta} | \mathbf{x})}{\mathcal{L}(\hat{\boldsymbol{\theta}} | \mathbf{x})} = 2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}) - 2 \log \mathcal{L}(\boldsymbol{\theta}) = q(\hat{\boldsymbol{\theta}}) \sim \chi_{(k)}^2$$

How far can get $\hat{\boldsymbol{\theta}}$ from $\boldsymbol{\theta}$ with CL% probability?

$$P \left[\left(2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}) - 2 \log \mathcal{L}(\boldsymbol{\theta}) \right) \leq c \right] = CL$$

Using Wilks' Theorem:

$$P \left[q(\hat{\boldsymbol{\theta}}) \leq c \right] = CL \Rightarrow c = \chi_{k,CL}^2 \text{ (CL quantile)}$$

$$P \left[\left(2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}) - 2 \log \mathcal{L}(\boldsymbol{\theta}) \right) \leq \chi_{k,CL}^2 \right] = CL$$

$$\log \mathcal{L}(\boldsymbol{\theta}) \geq \log \mathcal{L}(\hat{\boldsymbol{\theta}}) - \frac{1}{2} \chi_{k,CL}^2$$

LLR Confidence Intervals Theory: Multi-dimensional case

$$\log \mathcal{L}(\boldsymbol{\theta}) \geq \log \mathcal{L}(\hat{\boldsymbol{\theta}}) - \frac{1}{2} \chi_{k,CL}^2$$

If $f(x)$ has distribution $\chi_{(k)}^2$, the quantile $\chi_{k,CL}^2$ is defined as:

$$\int_0^{\chi_{k,CL}^2} f(x) \, dx = CL$$

LLR Confidence Intervals and nuisance parameters: Profile likelihood

Usually a measurement depends on several parameters, $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\theta} | \boldsymbol{x})$, divided into

- parameters of interest, $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_n\}$, usually one or two;
- the remaining ones, called nuisance parameters, $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$.

The **profile likelihood** is defined as:

$$\mathcal{L}_{\text{prof}}(\boldsymbol{\mu}) = \mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}}(\boldsymbol{\mu}))$$

where the notation $\hat{\boldsymbol{\theta}}(\boldsymbol{\mu})$ means the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}$, at fixed $\boldsymbol{\mu}$.

 Profiling produces a likelihood that depends only on the parameters of interest.

LLR Confidence Intervals and nuisance parameters: Profile likelihood

Wilks' theorem is valid for the profile likelihood ratio:

$$-2 \log \frac{\mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}})}{\mathcal{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})} = 2 \log \mathcal{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) - 2 \log \mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}}) \sim \chi_{(n)}^2$$

where n is the number of parameters of interest, $\boldsymbol{\mu}$.

This yields the same recipe for LLR confidence intervals when removing the nuisance parameters via profiling

$$\log \mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}}) = \mathcal{L}_{\max} - \frac{1}{2} \chi_{n, CL}^2$$

Lecture 2

Notebook time

- Notebooks:
 - Unbinned_Likelihood_fit_IDPASC.ipynb
 - Likelihood_IDPASC.ipynb
- Play with the others...

LLR Confidence Intervals Theory: Multi-dimensional case

$$\log \mathcal{L}(\boldsymbol{\theta}) \geq \log \mathcal{L}(\hat{\boldsymbol{\theta}}) - \frac{1}{2} \chi_{k,CL}^2$$

If $f(x)$ has distribution $\chi_{(k)}^2$, the quantile $\chi_{k,CL}^2$ is defined as:

$$\int_0^{\chi_{k,CL}^2} f(x) \, dx = CL$$

LLR Confidence Intervals and nuisance parameters: Profile likelihood

Usually a measurement depends on several parameters, $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\theta} | \boldsymbol{x})$, divided into

- parameters of interest, $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_n\}$, usually one or two;
- the remaining ones, called nuisance parameters, $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$.

The **profile likelihood** is defined as:

$$\mathcal{L}_{\text{prof}}(\boldsymbol{\mu}) = \mathcal{L}(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}}(\boldsymbol{\mu}))$$

where the notation $\hat{\boldsymbol{\theta}}(\boldsymbol{\mu})$ means the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}$, at fixed $\boldsymbol{\mu}$.

👉 Profiling produces a likelihood that depends only on the parameters of interest.

LLR Confidence Intervals and nuisance parameters: Profile likelihood

Wilks' theorem is valid for the profile likelihood ratio:

$$-2 \log \frac{\mathcal{L}(\mu, \hat{\theta})}{\mathcal{L}(\hat{\mu}, \hat{\theta})} = 2 \log \mathcal{L}(\hat{\mu}, \hat{\theta}) - 2 \log \mathcal{L}(\mu, \hat{\theta}) \sim \chi^2_{(n)}$$

where n is the number of parameters of interest, μ .

This yields the same recipe for LLR confidence intervals when removing the nuisance parameters via profiling

$$\log \mathcal{L}(\mu, \hat{\theta}) = \mathcal{L}_{\max} - \frac{1}{2} \chi^2_{n, CL}$$

A hint at Bayesian Statistics

Probability with more variables

- **Joint probability** $P(x,y)$

(probability of two things happening)

- If events are independent $P(x,y) = P(x) \times P(y)$

- **Marginal probability** $P(x)$

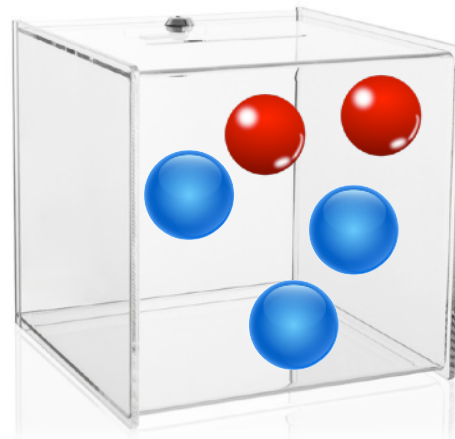
- Probability that x happens, regardless of y

- If discrete: $P(x) = \sum_y P(x, y)$

- Continuous case: $p(x) = \int p(x, y) dy$

Conditional Probability

- Probability that x happens, given that y happens (i.e. for a given value of y): $P(x|y)$
- If events are independent: $P(x|y) = P(x)$
- Not otherwise!
 - $P(\text{red})=0.4$
 - $P(\text{red}|\text{red})=0.25$
 - $P(\text{red}|\text{blue}) = 0.5$
- In general $P(x,y)=P(x|y)P(y)$



A worked example: disease in women and men

	Men	Women
Disease	200	100
No disease	300	400

$$P(X,Y) , X = [M,W], Y = [D, N]$$

Frequency yields (an estimate of!) probability

$$P(D,M) = 200/1000 = 0.2 \text{ (Joint probability)}$$



A worked example: disease in women and men

	Men	Women
Disease	200	100
No disease	300	400

$P(X,Y)$, $X = [M,W]$, $Y = [D, N]$

Frequency yields probability

$$P(D,M) = 200/1000 = 0.2$$



Marginal probabilities

- $P(D) = 300/1000 = 0.3$
- $P(M) = 500/1000 = 0.5$

Conditional probability

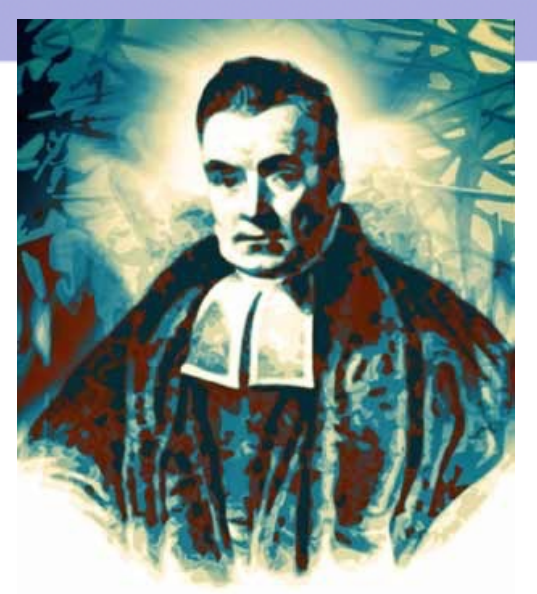
- $P(D|M) = 200/500 = 0.4$

Applying the rule

$$P(D|M) = \frac{P(D,M)}{P(M)}$$
$$= 0.2/0.5 = 0.4$$

(easily extended for PDFs!)

Bayesian Statistics



- Frequentist probability: rates at which events occur (needs many repetitions!)
- Bayesian probability: degree of belief
- Powerful approach: inference, confidence, use prior information, updates
- Widely used in physics and, more recently, in AI
- Bayes theorem:
 - From the conditional and marginal probabilities

$$P(X, Y) = P(X|Y) \times P(Y)$$

$$P(X, Y) = P(Y|X) \times P(X)$$

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Example: am I infected?

- Test for a disease is 90% sensitive: true positive rate (TPP) = 0.9
- It is 80% specific: True Negative Rate (TNR) = 0.8 (FPR = 0.2)
- Prevalence (fraction of infected): 5%
- So, if I test positive, means there is 90% chance I am infected?

$$P(\text{Infected}|\text{Positive}) = \frac{P(\text{Positive}|\text{Infected})P(\text{Infected})}{P(\text{Positive})}$$

$$= \frac{P(\text{Positive}|\text{Infected})P(\text{Infected})}{P(\text{Positive}|\text{Infected})P(\text{Infected}) + P(\text{Positive}|\text{NotInfected})P(\text{NotInfected})}$$

$$= \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.2 \times 0.95} \simeq 19\%$$

Modelling data

- Bayesian statistics is extremely powerful for modeling data

posterior: $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$

model (points to $P(X|Y)$)
likelihood (points to $P(Y|X)$)
prior (points to $P(X)$)
data (points to Y in $P(X|Y)$)

In practice $P(X|Y) \propto P(Y|X)P(X)$

Constant of proportionality derived from normalization (probabilities add to 1)

- Maximizing the posterior with respect to X one obtains the best model (within a family) that describes the data
- Yields the whole pdf of X : shape, confidence levels / error bars....

Bayesian statistics for continuous variables

- The likelihood is the joint probability density function (PDF) of the observed data, viewed as a function of the parameter vector

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x} \mid \boldsymbol{\theta})$$

- For n independent and identically distributed observations:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i \mid \boldsymbol{\theta})$$

- New ingredient: prior Distribution $\pi(\boldsymbol{\theta})$
- A joint probability density function representing prior information:

$$\int_{\Theta} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$$

Bayes theorem

- To combine the likelihood and the prior into a valid probability density function, we normalize by the marginal likelihood (also called the evidence):

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{p(\mathbf{x})}$$

- Where the marginal likelihood is the multi-dimensional definite integral:

$$p(\mathbf{x}) = \int_{\Theta} f(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\Theta_1} \cdots \int_{\Theta_d} f(\mathbf{x} \mid \theta_1, \dots, \theta_d) \pi(\theta_1, \dots, \theta_d) d\theta_1 \dots d\theta_d$$

- Require sophisticated numerical methods (e.g. MCMC)

Marginal Posterior Density

- If you only care about a single parameter and want to integrate out the nuisance parameters

$$p(\theta_1 \mid \mathbf{x}) = \int_{\Theta_2} \cdots \int_{\Theta_d} p(\theta_1, \theta_2, \dots, \theta_d \mid \mathbf{x}) d\theta_2 \dots d\theta_d$$

Bayesian Statistics

- Bayesian methods are extremely powerful
 - inference
 - use prior information
 - confidence
 - can learn from new data: the output (Posterior) becomes your new Prior when new data comes in
 - in Machine Learning: regression, error bars
- Real world applications use all concepts of this class

Least Squares and Model Fitting

Maximum Likelihood Principle $\rightarrow$ Least Squares Method

For n measurements from a distribution $f(x | \theta)$, the likelihood is $\mathcal{L}(\theta) = \prod f(x_i | \theta)$.

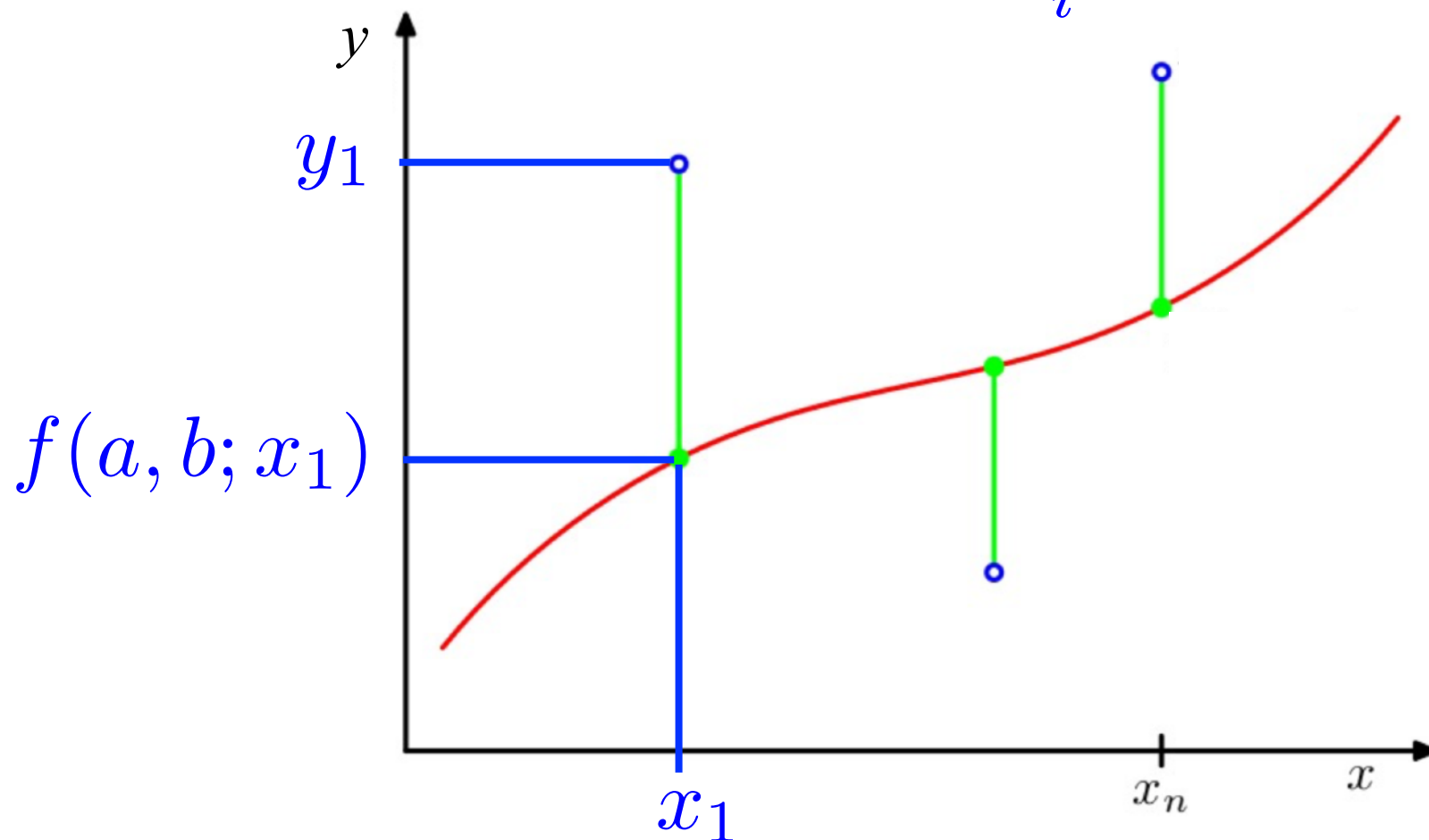
When $f(x | \theta)$ is Gaussian, the ML principle yields the method of least squares,
also called χ^2 minimization:

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(x_i | \theta) = C \prod_{i=1}^n e^{-\frac{1}{2} \left(\frac{x_i - \mu(\theta)}{\sigma_i} \right)^2} \Rightarrow -2 \log \mathcal{L}(\theta) = \sum_{i=1}^n \left(\frac{x_i - \mu(\theta)}{\sigma_i} \right)^2 + C'$$

Least squares regression

- Find a model (function) that represents (approximately) the data
- Minimize the distance of a model to the data

$$D^2(a, b; \{y\}, \{x\}) = \sum_i (f(a, b; x_i) - y_i)^2$$



Goodness of the fit

- In a normal distribution

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i | \boldsymbol{\theta}) = C \prod_{i=1}^n e^{-\frac{1}{2} \left(\frac{x_i - \mu(\boldsymbol{\theta})}{\sigma_i} \right)^2}$$

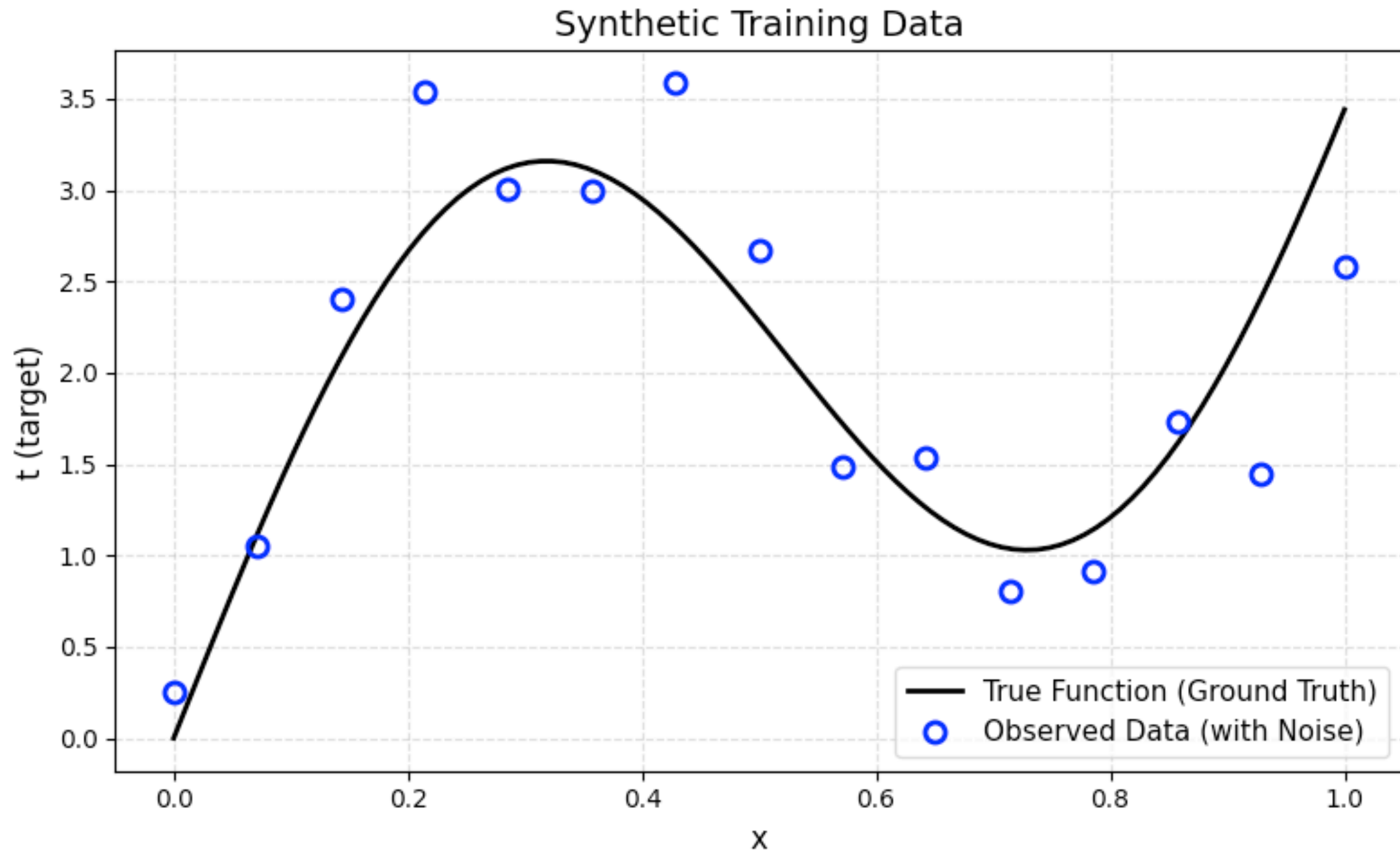
by definition, the random variable $Z_i = \frac{x_i - \mu_i(\theta)}{\sigma_i}$

becomes a Standard Gaussian variable with a mean of 0 and a variance of 1 so we expect, for $N-k$ degrees of freedom (dof):

- $\chi^2/\text{dof} \approx 1$: The data points deviate from the model by exactly the amount expected from their 1σ measurement errors.
- $\chi^2/\text{dof} \gg 1$: The model is a poor fit (underfitting), or you have underestimated your measurement errors.
- $\chi^2/\text{dof} \ll 1$: The model is tracking the noise instead of the physics (overfitting), or you have overestimated your measurement errors.

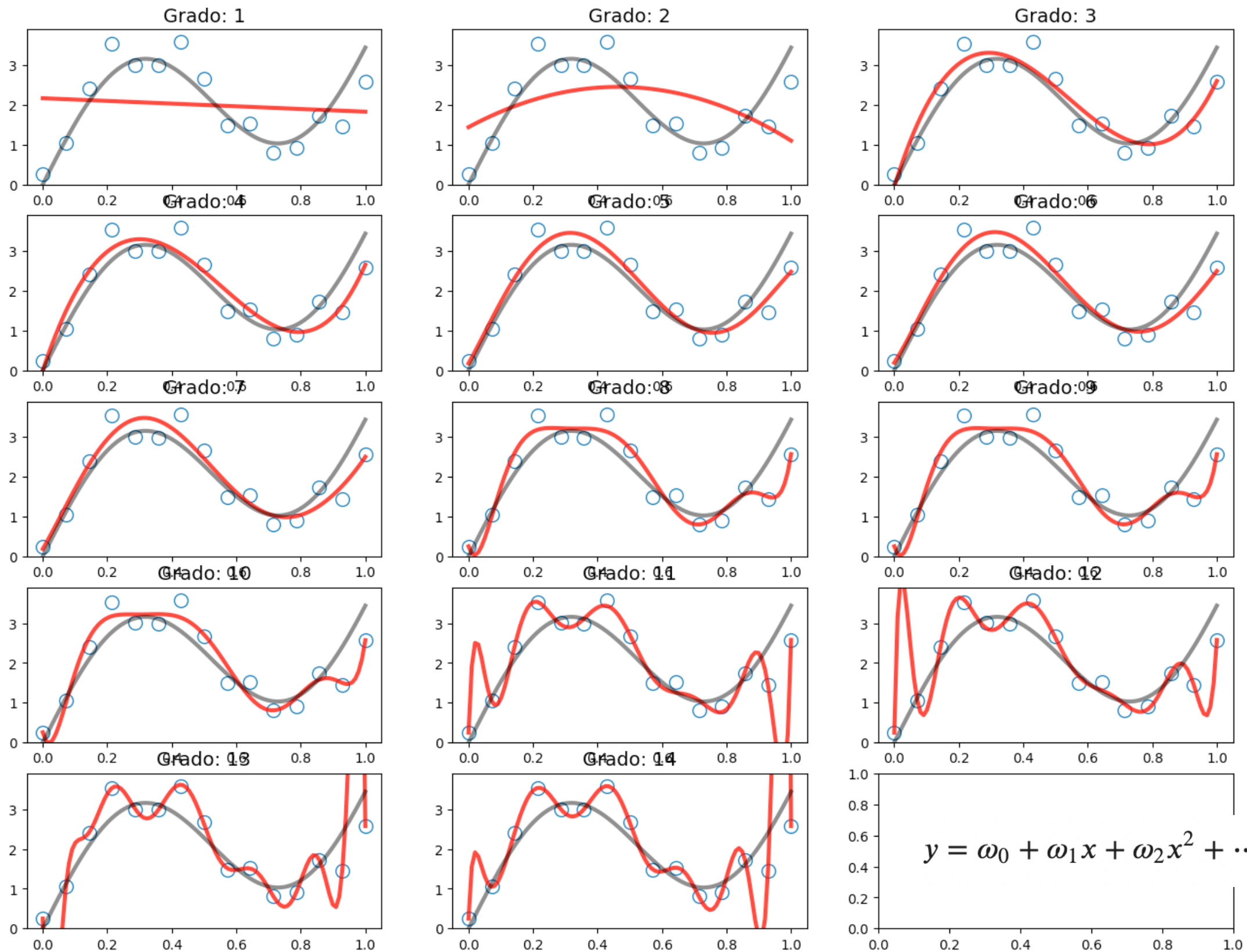
Over and Under-fitting

Play with notebook: PolynomialRegression.ipynb

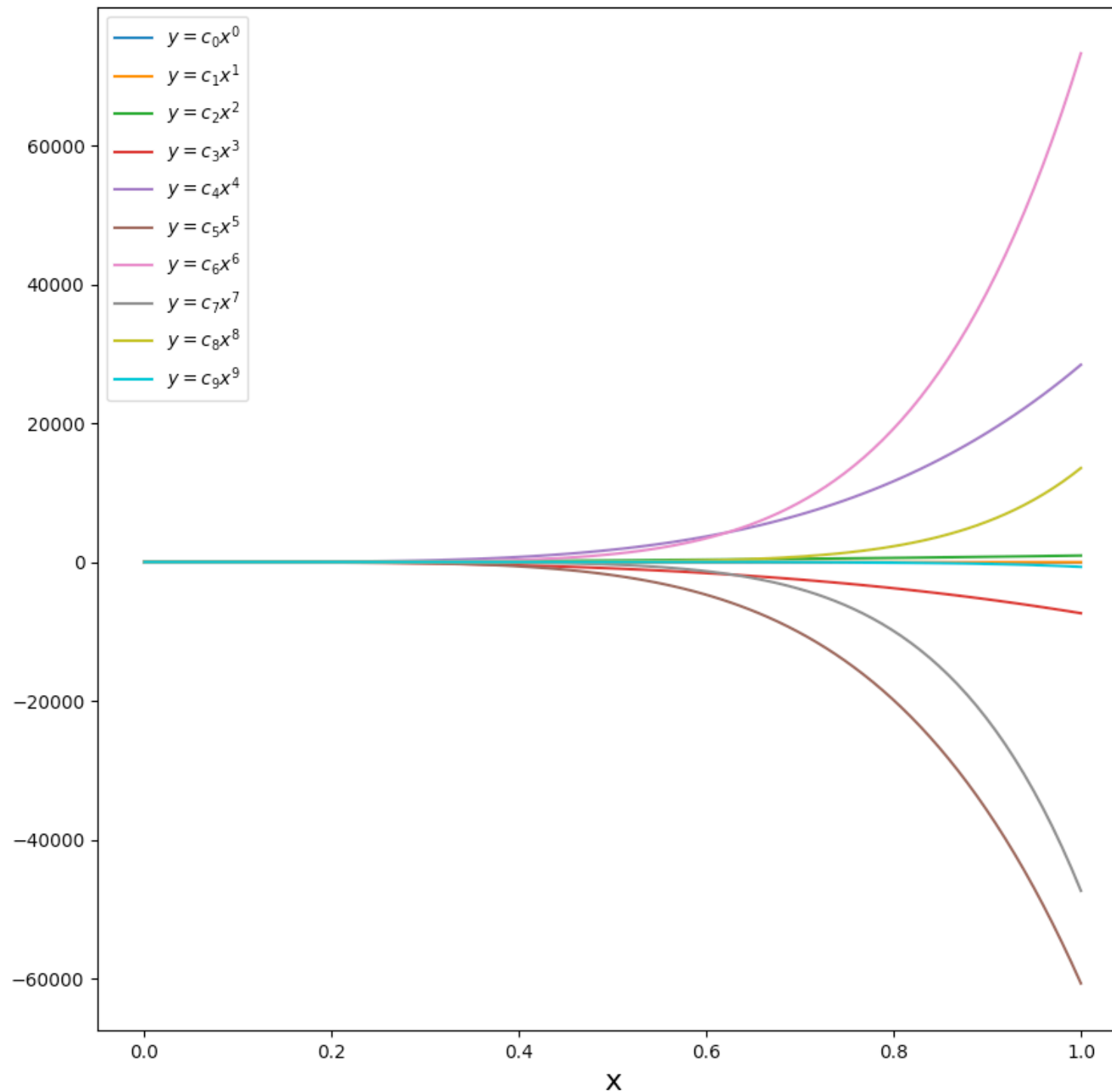


$$y = \omega_0 + \omega_1 x + \omega_2 x^2 + \cdots + \omega_M x^M$$

Increasing the degree



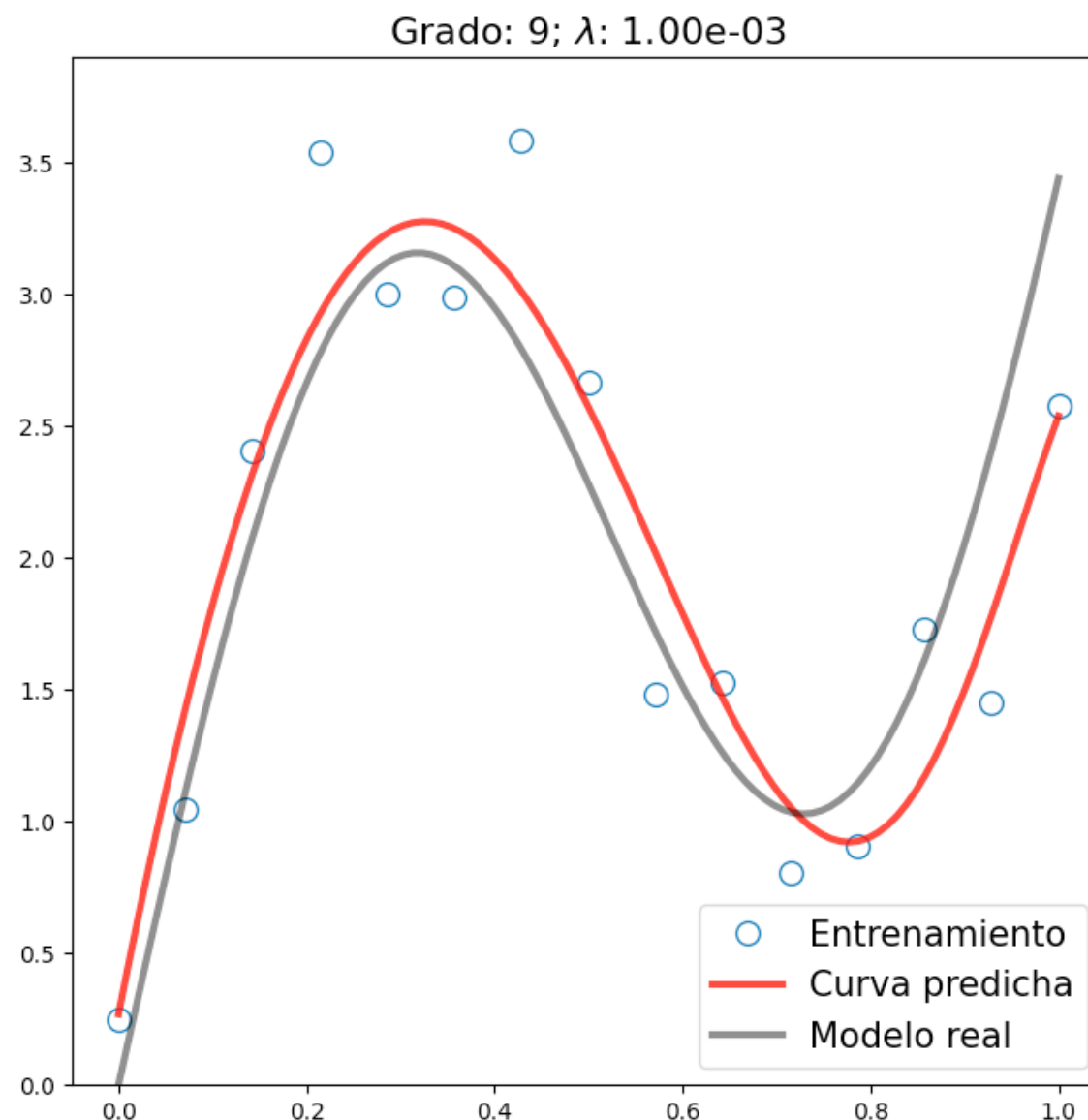
Overfitting: coefficients diverge



Empirical: regularization

- Solution: add a regularization term

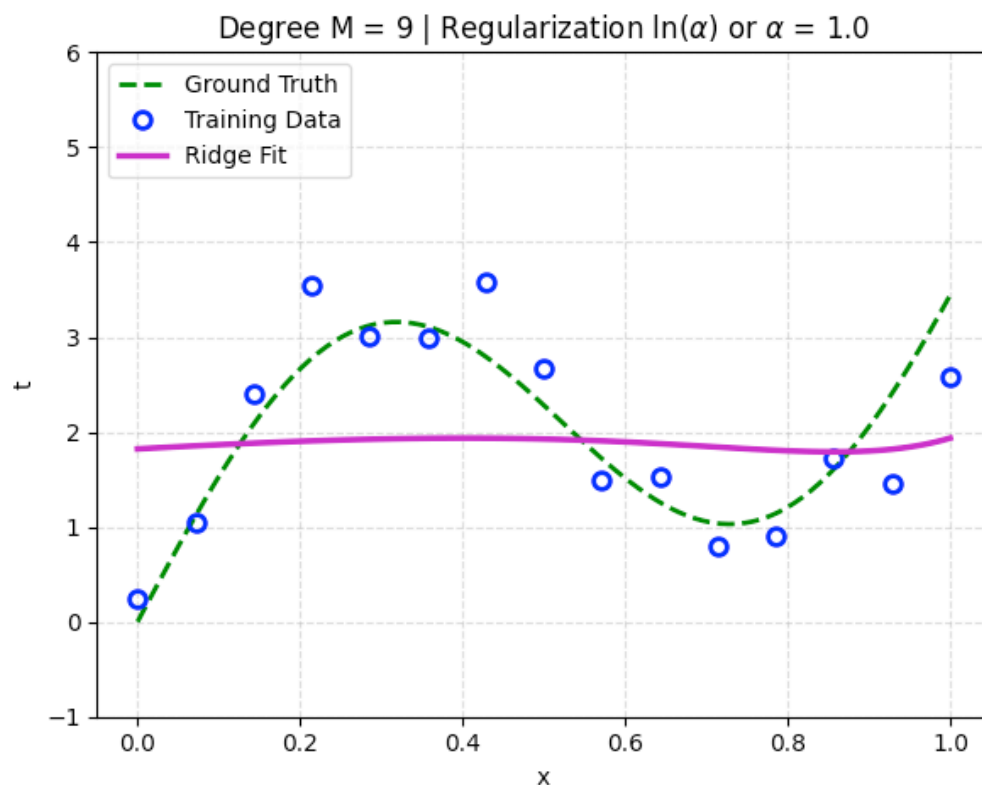
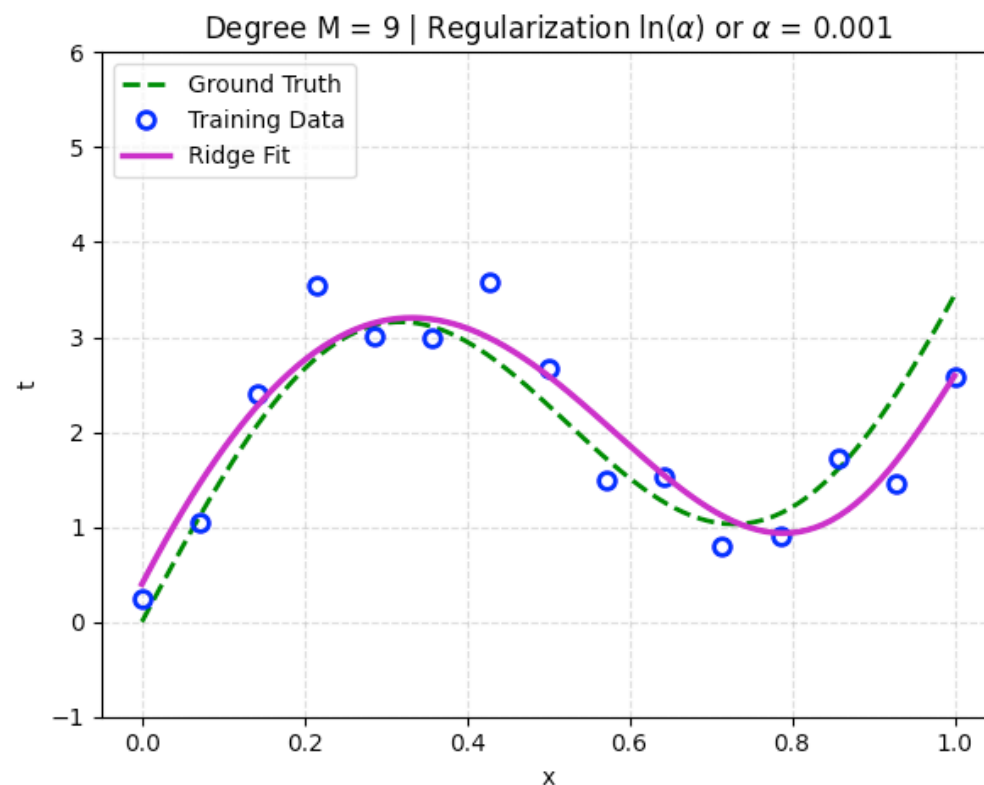
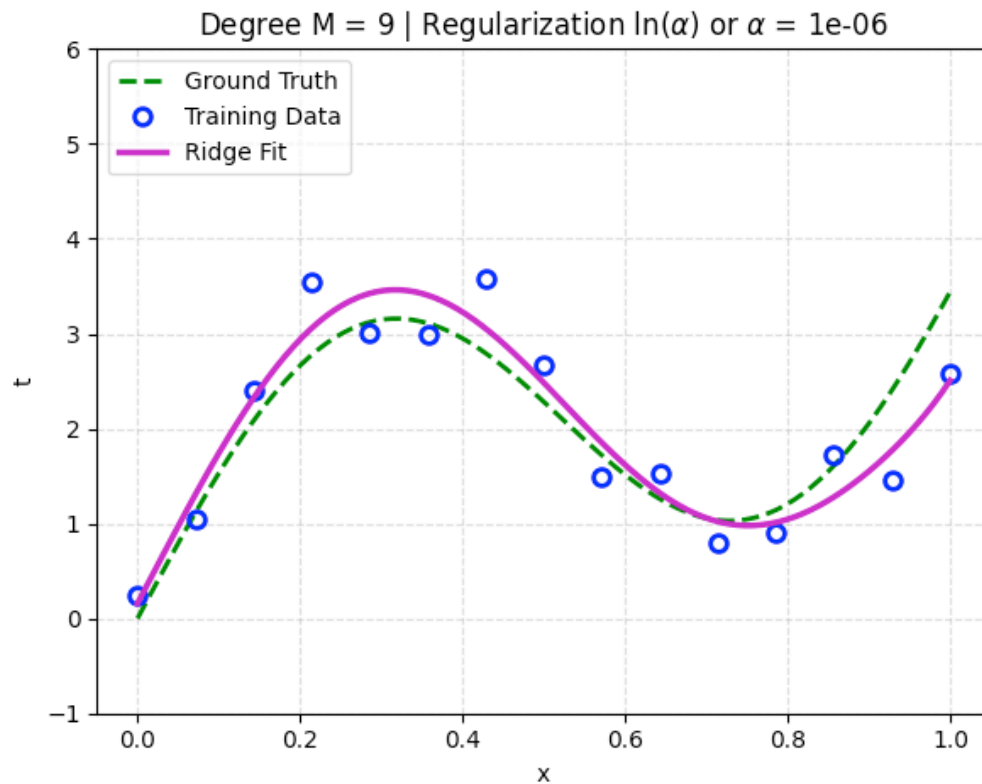
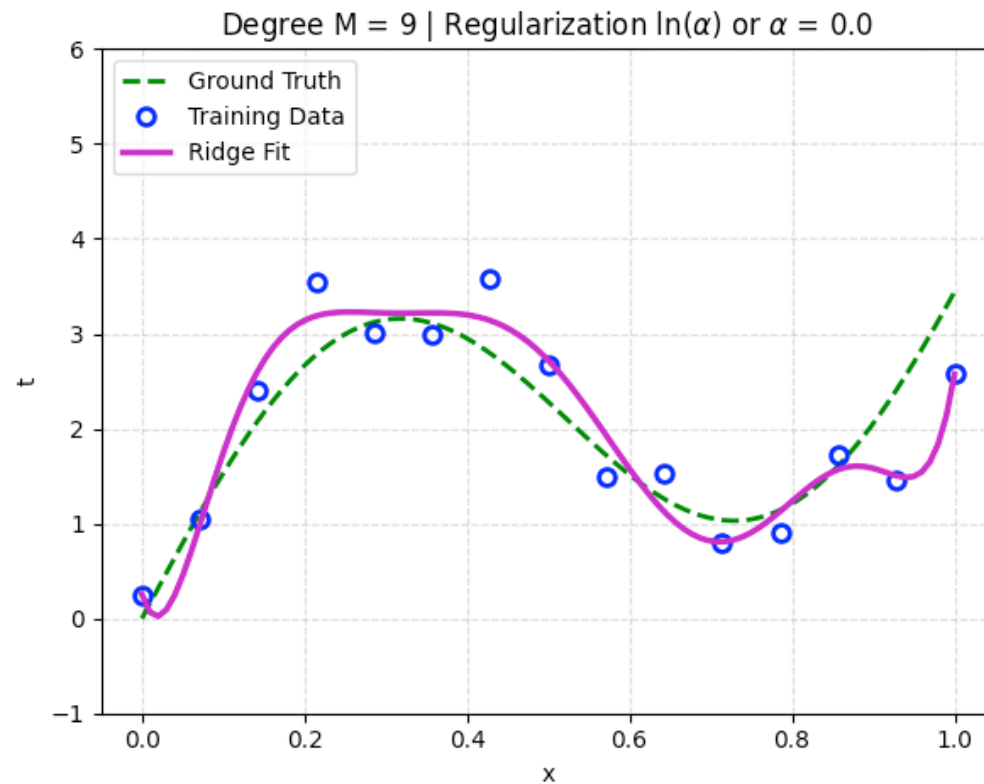
$$E_{\text{ridge}}(\boldsymbol{\omega}; \lambda) = \frac{1}{2} \sum_{i=1}^N \{y(x_i, \boldsymbol{\omega}) - t_i\}^2 + \frac{\lambda}{2} \sum_{i=1}^M \omega_i^2$$



Avoids overfitting

Tuning the regularization parameter

Impact of L2 Regularization on Overfitted Model ($M = 9$)



Physics: back to statistics!

- Goodness of fit
- Bayesian evidence—also called the marginal likelihood—is the normalizing constant in Bayes' theorem. It integrates the product of the likelihood

$$\mathcal{Z} \equiv P(d \mid M) = \int_{\Omega} \mathcal{L}(\vec{\theta}) \pi(\vec{\theta}) d^D \vec{\theta}$$

- Unlike a maximum-likelihood fit or χ^2 , which only looks at the absolute best-fit coordinate, the evidence calculates the average performance of the model across its entire prior range.
- “Built-in” Occam’s factor
- Comparing models: evidence ratio, Jeffreys Scale

What we touched upon

- Likelihood from probability (and their differences)
- Maximum likelihood method
- Uncertainties on the model parameters from likelihood ratio (Gaussian or not)
- Profile likelihood for nuisance parameters
- A peek at Bayesian statistics
- A hint at model fitting

What we didn't address

- Hypothesis Testing & Discovery Criteria
- Goodness-of-Fit & Model Comparison
- Bootstrap - resampling
- Systematic errors
- Connected to Bayesian Statistics & Inference
 - Markov Chain Monte Carlo (MCMC): Algorithms for marginalizing over high-dimensional parameter spaces
 - Nested Sampling: Algorithms like MultiNest or Dynesty used for calculating Bayesian evidence and model selection in cosmology.
- Modern methods connected to Deep Learning
 - Neural posterior estimators
 - Simulation based inference
- Etc., etc.

Discussion

- Statistics is a key element in physics, and specially in particle physics, astrophysics and cosmology
- Introduced a few key concepts and worked simple examples in more detail, but covered a tiny part of the subject
- We are in the era of Big Data and Deep Learning. Modern statistical methods are an important part, but never give up the understanding!
- Use and have fun with the notebooks

Credits

Part 1 of these lectures was taken from the lectures *Confidence intervals in particle physics* by Dario Rodrigues (University of Buenos Aires - UBA), presented at the Magnificent CEvNS School at CBPF, Rio de Janeiro (<https://indico.cern.ch/event/1458241/>), which were originally prepared by Professor Ricardo Piegaia from UBA.